Novel Statistical Methods Applied in Clinical Trials and Gut Microbiota

Dissertation Advisor: Professor Marcello Pagano Richard Aubrey White

Novel Statistical Methods Applied in Clinical Trials and Gut Microbiota

Abstract

Ethical clinical trials need both societal and personal equipoise. Recently, personal equipoise has been disturbed by the introduction of interim analyses; after an interim analysis has been performed the study administrators have additional information about the treatments, which is withheld from new recruits. For true informed consent, this information should be given to new study recruits to use in making a personal decision about their desired treatment. We present a method (and the rationale behind the method) that provides unbiased estimates of hazard ratios when new recruits are given information from interim analyses and allowed to choose their own treatments. We then developed a novel procedure that allows for the identification of longitudinal gut microbiota patterns (corresponding to the gut ecosystem evolving), which are associated with an outcome of interest, while appropriately controlling for the false discovery rate. Finally, using novel statistical models, we investigated the impact of POPs (in particular, non-dioxin-like polychlorinated biphenyl, IUPAC no.: 153; "PCB153") on human health through the disruption of natural gut microbiota establishment in infants. We created novel distributed lag two-part models to account for the cumulative exposure of POPs.

Contents

Acknowledgements

This work was a joint collaboration with the Norwegian Institute of Public Health, Division of Epidemiology, Department of Environment and Genes (EPAM). Dearest and most grateful thanks for all of its support, both intellectually and financially.

Obvious thanks go to Marcello Pagano, a fantastic supervisor, and all around great person. Similar shout-outs to Winston Hide, Francesca Dominici, and Merete Eggesbø.

More specific thanks (here not limited to just the NIPH) go to Merete Eggesbø (again!), Shyamal Peddada, Jørgen Bjørnholt, Tore Midtvedt, Donna Baird, Carole Mitnick, Matt Miller, Deb Azrael, and Cathy Barber.

Of course there are others who bear mentioning, such as Sam E, for restoring my faith that man is once again capable of greatness. Before him, I only had Prometheus to look to for inspiration. To quote him: "I mingle with my peers or no one, and since I have no peers, I mingle with no one."

Emmanuel joined the senselessness of a Caligula with the refinement of a Borgia. He was as good at shooting crap as he was at playing Ping-Pong, and he was as good at playing Ping-Pong as he was at everything else. Everything Emmanuel did, he did well. He was a fair-haired boy from Cyprus who believed in God, Motherhood, and the American Way of Life, without ever thinking about any of them, and everybody who knew him liked him.

Chris was not only the Vice-Shah or Oran, as it turned out, but also the Caliph of Baghdad, the Imam of Damascus, and the Sheik of Araby. Chris was the corn god, the rain god and the rice god in backward regions where such crude gods were still worshipped by ignorant and superstitious people, and deep inside the jungles of Africa, he intimated with becoming modesty, large graven images of his mustached face could be found overlooking primitive stone altars red with human blood. Everywhere he was acclaimed with honour, and it was one triumphal ovation after another for him in city after city.

And in memory of teaching at Harvard: "At the height of the demonstration I dumped all of the old papers - ungraded, of course - out of the window and right onto the students' heads.

The college was too small to accept this act of defiance against the abyss of contemporary academia. I also told the students that, for the sake of humanity's future, I hoped they were all sterile. I could never have possibly read over the illiteracies and misconceptions burbling from the dark minds of these students."

Chapter 1

Introduction

This thesis is constructed from three papers; the first concerns ethics in clinical trials, and the later two are focused on gut microbiota. The gut microbiota papers were produced while working at the Norwegian Institute of Public Health.

In the first paper, we discuss that ethical clinical trials need both societal and personal equipoise. Recently, personal equipoise has been disturbed by the introduction of interim analyses; after an interim analysis has been performed the study administrators have additional information about the treatments, which is withheld from new recruits. For true informed consent, this information should be given to new study recruits to use in making a personal decision about their desired treatment. We present a method (and the rationale behind the method) that provides unbiased estimates of hazard ratios when new recruits are given information from interim analyses and allowed to choose their own treatments.

We studied the concept of clinical trials in general, and used multiple simulated clinical trials to demonstrate our method. The main intervention of interest was allowing recruited subjects to choose their own treatments, after informing them of the interim trial results. We assessed our methods effectiveness by observing the bias of the estimated hazard ratio in our simulated clinical trials, and when the true hazard ratio lay between 0.65 and 1, our method's bias was within ± 0.01. We demonstrate that our method could have saved six

women from breast cancer and ten men from HIV infection, in two historical clinical trials. The main limitation of our method is the risk of bias if confounders are misspecified in the model.

True informed consent requires transparency; that is, current information. If the trial has not been terminated early, then information from the interim analysis must be used to update the informed consent. This is the first paper to propose and validate a random clinical trial protocol that would allow subjects to choose their own treatments, while still producing unbiased hazard ratio estimates that can be used in hypothesis tests.

The second paper revolves around obesity and gut microbiota in infants. It is widely acknowledged that obesity trajectories are set early in life, and that rapid weight gain in infancy is a risk factor for later development of obesity. Identifying modifiable factors associated with early rapid weight gain is a prerequisite for curtailing the growing worldwide obesity epidemic. Recently, much attention has been given to findings indicating that gut microbiota may play a role in obesity development. This is the first longitudinal study that aims at identifying how the development of early gut microbiota is associated with expected infant growth. We developed a novel procedure that allows for the identification of longitudinal gut microbiota patterns (corresponding to the gut ecosystem evolving), which are associated with an outcome of interest, while appropriately controlling for the false discovery rate. Our method identified developmental pathways of Staphylococcus species and Escherichia coli that were associated with expected growth. Our method should have wide future applicability for studying gut microbiota, and is particularly important for translational considerations, as it is critical to understand the correct timing and design of the interventions, prior to attempting to manipulate gut microbiota in early life.

The final paper is concerned with how PCB153 affects gut microbiota in early infancy. Gut microbiota has a critical role in human health; understanding its role in early infancy is of particular interest due to the time dependent windows that rely on microbial stimulus from the gut. That is, the development of tolerance and the optimal functioning of angiogen-

esis and stress responses later in life, require time dependent actions in the gut. Persistent organic pollutants (POPs) are widespread environmental contaminants that are resistant to environmental degradation through normal processes, which causes them to bioaccumulate in human and animal tissue and biomagnify in food chains. POPs are known carcinogens that disrupt natural human systems (endocrine, reproductive, and immune). Using novel statistical models, we investigated the impact of POPs (in particular, non-dioxin-like polychlorinated biphenyl, IUPAC no.: 153; "PCB153") on human health through the disruption of natural gut microbiota establishment in infants. We created novel distributed lag two-part models to account for the cumulative exposure of POPs. We then identified significant associations concerning POPs affecting gut microbiota species (spp.) groups (from birth through to day 120 of life). Strong associations were found between POPs and *Bifidobacterium* spp., *Bifidobacterium bifidum*, and *Lactobacillus* spp.. Using these findings we successfully identified gut microbiota as a potential vector through which POPs may harm humans; examples were given for POPs acting as carcinogens and diarrhoeal agents.

Chapter 2

Subjects in randomized clinical trials should be allowed to choose their own treatments

2.1 Abstract

2.1.1 Background

Ethical clinical trials need both societal and personal equipoise. Recently, personal equipoise has been disturbed by the introduction of interim analyses; after an interim analysis has been performed the study administrators have additional information about the treatments, which is withheld from new recruits. For true informed consent, this information should be given to new study recruits to use in making a personal decision about their desired treatment. We present a method (and the rationale behind the method) that provides unbiased estimates of hazard ratios when new recruits are given information from interim analyses and allowed to choose their own treatments.

2.1.2 Methods and findings

We studied the concept of clinical trials in general, and used multiple simulated clinical trials to demonstrate our method. The main intervention of interest was allowing recruited subjects to choose their own treatments, after informing them of the interim trial results. We assessed our methods effectiveness by observing the bias of the estimated hazard ratio in our simulated clinical trials, and when the true hazard ratio lay between 0.65 and 1, our method's bias was within ± 0.01. We demonstrate that our method could have saved six women from breast cancer and ten men from HIV infection, in two historical clinical trials. The main limitation of our method is the risk of bias if confounders are misspecified in the model.

2.1.3 Conclusion

True informed consent requires transparency; that is, current information. If the trial has not been terminated early, then information from the interim analysis must be used to update the informed consent. This is the first paper to propose and validate a random clinical trial protocol that would allow subjects to choose their own treatments, while still producing unbiased hazard ratio estimates that can be used in hypothesis tests.

2.2 Introduction

The randomized clinical trial has been called one of the "ten definitive moments" in medical advances of the twentieth century. It has evolved into "the gold standard by which the merits of modern drug therapy must be measured."[1] In this evolution care has also been taken to produce an ethical procedure, but as sometimes happens when building a multi-story building in stages, one needs to recheck the foundations as the building gets taller, so too with the randomized clinical trial, we must periodically question our procedures. Of

[1] J LeFanu: The Rise and Fall of Modern Medicine, New York 2000.

late we have started requiring interim analyses in clinical trials, and this may have upset a delicate balance. The time has come for us to expend some thought on how to proceed ethically.

Clinical trials have an interesting history. Briefly looking back in time we find:

> *For Medicine is not a naked word, a vain boasting, or vain talk, for it leaves a work behind it: Wherefore I despise reproaches, the boastings, and miserable vanities of ambition: Go to return with me to the purpose: If ye speak truth, Oh ye Schooles, that ye can cure any kinde of Fevers without evacuation, but will not for fear of a worse relapse, come down to the contest ye Humorists: Let us take out of the Hospitals, out of the Camps, or from elsewhere, 200, or 500 poor People, that have Fevers, Pleurisies, etc. Let us divide them in halfes, let us cast lots, that one halfe of them may fall to my share, and the other to yours; I will cure them without blood-letting and sensible evacuation; but do you do, as ye know (for neither do I tye you up to the boasting, or of Phlebotomy, or the abstinence from a solutive Medicine) we shall see how many Funerals both of us shall have: But let the reward of the contention or wager, be 300 Florens, deposited on both sides: Here your business is decided.*

And so the gauntlet was flung down by the great physician Jean Baptiste van Helmont (1577-1644) in order to prove his point of view[2]. Progress can only be made by exploring the unknown, by experimentation. With humans we have options, we can reason; we can extrapolate from the laboratory; the animal model; and even use simulation models on the computer, but, as has been shown repeatedly, the final proof always must come from the clinical trial.

This point was clear to the townspeople of Milton, Massachusetts who in 1809 vaccinated

[2]J vanHelmont: Oriatrike, or, Physick refined: the common errors therein refuted, and the whole art reformed & rectified: being a new rise and progress of philosophy and medicine for the destruction of diseases and prolongation of life now faithfully rendered into English by J.C., sometime of M.H. Oxon.

a quarter of the town's inhabitants with the Kine Pox[3]. In order to document the effectiveness of the procedure, they invited dignitaries, including the health commissioner from Boston, to witness a clinical trial. They enlisted twelve children who had been vaccinated (and their names are given in the report; this was pre-HIPAA). Each of these children was administered some dry smallpox the town had acquired. They were then placed in an isolated house, and after fifteen days released to prove to the world that they were free of the smallpox.

Presumably there was no institutional review board in Milton at the time, although why the parents or guardians of these children would allow such an experiment to take place is astounding to our current sensibilities. There is no mention of how these children were chosen. In van Helmont's case, on the other hand, it is noteworthy to a statistician that to make the contest fair, and presumably attractive to his adversaries, he proposed casting lots to see who got to treat which half. Not attractive enough, as there is no record of anyone taking him up on his challenge, but an accepted way, even at that time, to balance the unknown. Indeed, this is not the first time in history we see lots being used as arbitrage instruments; after all, the Roman soldiers at the foot of the cross in Calgary, rather than tear Christ's garment into equal sized pieces, apparently drew lots to see who would get the whole vesture.

This quest to make the unpredictable at least seem fair, has been the basis for gambling presumably from time immemorial. The desired end may be achieved by the use of a randomization device such as an astralagus[4], or a die, or today, the computer; although to be precise the computer generated numbers we use in modern random patient assignments are not random but rather pseudo random numbers, although these retain the important features of random numbers. It may not be pure coincidence that the first book written on probability was by the famous Milanese physician Girolamo Cardano (Hieronimus Cardanus, to give him his Latin name)(1501-1576)[5], and indeed it must have been his passion

[3]Milton.

[4]F David: Games, Gods, & Gambling: A History of Probability and Statistical Ideas.

[5]O Ore: Cardano: The Gambling Scholar; With a Translation from the Latin of Cardano's 'Book on Games of Chance,' by Sydney Henry Gould.

(and necessity, the Milanese guild would not let him practise medicine in Milan) for gambling that drove him to the study of probability, but the poor predictability of the outcomes of medical treatments in his day may also have influenced his decision to pursue his study of probability.

It is interesting that a random device was to play a role in the early, proposed clinical trial of van Helmont. That the practice seemed desirable is also evident in the trials in the early part of the 20th century where we see that balance in treatment assignment was maintained by alternating the patients between treatments, the order being determined by when they arrived on trial. This is just as good a random device as any–arguing that when a patient shows up is a random event–except that it had a component of predictability, and thus its intent could be easily subverted; by holding back a patient, for example. Today we sometimes use what are called blocked randomization schemes[6] where, in multi-center trials especially, a block size is chosen, such as two or four, so that after each block of patients the treatment allocations are balanced. This is not too far from the alternating scheme mentioned above, but is a little less predictable.

The desirability of having unpredictable treatment assignment led to the modern randomized trial. The most famous early one being the streptomycin tuberculosis trial in the UK where, because of the shortage of the availability of the treatment, investigators were led to draw lots to see who got the streptomycin. As a result they were then spared from deciding directly who was assigned to which treatment. There was one exception, of course, when one of the physicians involved was afflicted by tuberculosis; he was given the drug.

The current rationale for randomization has not changed and still remains one of balance, especially balance between unknown factors. We follow Laplace's dictum, "Probability is the reflection of man's ignorance," when we appeal to randomization to balance what we cannot. This is our attempt in clinical trials to achieve the equality available to physical scientists: If two identical physical quantities are treated differently and subsequently found to differ on

[6]M Zelen: The randomization and stratification of patients to clinical trials. In: Journal Chronic Diseases 27 (1974), pp. 365–375.

some measure, then the difference is attributed to the different treatments. This is logically sound because the two started off as being identical. The problem with humans is that, even if we have an ethical trial design, there are not two identical individuals. So to control this inherent variability, we create two groups that are as balanced as we can make them, so that if we treat the two groups differently and do find a difference between them, then we can logically attribute the difference in outcome to the difference in treatment. That is the theory.

In practice, just as there are no two identical individuals, so too we cannot find two groups perfectly balanced. But if we use a randomization device, then "on average" the two groups are balanced. Since we follow Laplace's pronouncement, we force balance according to some criterion or criteria, such as age, sex, etc., sometimes called confounders, that we know need to be balanced in order to yield a sensible result when we compare the two groups. Note that the more of these confounders we force balance on, the more predictable the assignment becomes, so we tend to balance on few of these and rely on chance to provide the balance, on average, and then use statistical models to account for the rest. If we ignore an important confounder, such as sex may be on a particular study, the random assignment may result in all women on one treatment and all men on another, and no statistical model can rescue us then, but that is quite unlikely to happen, or at least that is what we believe. The same logic holds for unknown factors.

This random allocation in a randomized clinical trial is one way of achieving fairness, and it also allows the statistician to use probabilistic methods to analyze the trial. Underlying the acceptance of this randomization is the belief that both treatments are equally good. And this, of course, is the crux of a hotly debated topic. How can two treatments be equally good? This is where the notion of equipoise, first formulated by Freedman[7], is usually introduced. If we phrase it in the case of the generic experimental-versus-standard-treatment framework, we have that equipoise is the belief, by each investigator who places a patient on the study,

[7]B Freedman: Equipoise and the ethics of clinical research. In: New England Journal of Medicine 317 (1987), pp. 141–145.

that the two treatments are equally good. To ease this maybe unrealistic requirement, we have the interpretation that "the society of knowledgeable practitioners" believes that the two are equally effective.[8]

That not everyone believes this equivalence between treatments is clear. For example, if the trial is being privately funded then some group has decided to spend a lot of time and money on an experimental treatment that surely they think has a good chance of being proved better, or at least as good as the standard treatment. But such investors have been proven wrong so often that reasonable people may discount their predictive capabilities.

But equipoise is a way of achieving a balance between two, at times conflicting, wishes: (a) those of the individual, and presumably his or her physician, who wishes the best available treatment, and (b) physicians' and society's desire to gain knowledge in order to better treat future patients. This belief, together with an informed consent of the participants, or their proxies, is the foundation for an ethical trial.

With equipoise reached, one decides the size of the trial to run. Van Helmont felt he needed "200 or 500 poor people" to make his point, whereas the townspeople of Milton used only a dozen boys. The size of the trial is important. It is unethical to have too small a trial just as it is to have too large a trial. When comparing two treatments, say, if the trial is too small, or underpowered, then chances are that no difference will be found between the two treatments. That may be the undeclared intent of the trial, but then why subject patients to any risk? If one treatment is indeed better than the other, we would wish to find this out as soon as possible and place as small a number of patients on the inferior treatment as possible. In summary, the number of patients one needs on a trial is related to how much trust we need to place on the results of the trial.

Once the trial is underway, then the delicate balance at the beginning of a trial is increasingly threatened as more knowledge is gained. It is certainly lost if one treatment is declared better than the other, as most published clinical trials decide. The nub is that

[8]R Lilford: Equipoise and the ethics of randomization, in: Journal of the Royal Society of Medicine 88 (1995), pp. 552–559.

to make such a pronouncement for future–untold numbers of–patients may require a higher level of confidence than if one needs to make the decision just at the personal level.

This point becomes decisive after performing interim analyses in a clinical trial; especially in sequential trials. The data and safety monitoring committee that oversees clinical trials is usually charged with making sure that interim analyses are carried out periodically as the trial proceeds[9]. At each of the meetings of this committee a decision is made whether to continue or terminate the trial. The latter option is made for various reasons: for example, safety (the experimental treatment is too toxic); or one treatment has been shown to be preferred to the other to the desired degree of confidence; or futility (it does not matter how the rest of the trial goes, the current preference-decision will not be changed).

These are all laudable aims, and this committee is a great improvement in the running of clinical trials, but what happens at each one of these meetings is that we get progressively more and more information about the two treatments. At each meeting the information is usually kept secret *if the trial is to continue*, but the information should be understandable to anyone who can interpret statistics-speak. But the problem is that equipoise may now be disturbed by the new information, even if the decision by the committee is to continue the trial. This then raises the question of whether it is still ethical to ask more patients to come onto a trial whilst hiding this newly acquired information from them.

Of course, the crux of the argument is if we think that equipoise has been disturbed. But how do we determine that? For whom has equipoise been disturbed? This is where the next patient-to-come-on-trial's aims and society's aims possibly become diametrically opposed. As far as societal benefit being best served, the answer is to continue the trial since statistical significance has not been achieved and thus to the certainty required (however we measure it) for a societal decision to be made, equipoise has not been disturbed. But when it comes to the individual patient who may not require as much convincing to believe that one treatment is better than the other, the answer is not clear; equipoise may thus well be

[9]J Wittes: Forming your phase iii trial's data and safety monitoring board: A perspective on safety, in: Journal of Investigative Medicine 52 (2004), p. 7.

broached for some individuals, although not for society.

On a personal level, which one of us would not prefer treatment A over treatment B where the two are comparable inter alia (toxicity, cost, length of treatment etc.) and our best estimate so far is that A has a 0.6 probability of success and B has a 0.4 probability of success? It may well be that these estimated probabilities come with standard errors of 0.2 so the differences are not statistically significant at a reasonable p-value, and as a result if we consider making a recommendation for all future patients, we need to get more information. But the question is when one is making a personal choice for oneself or a loved one, which would one choose? Society may be correct in not wishing to make any pronouncement about the two treatments being significantly different, but an intelligent individual may well think them sufficiently different to make an individual choice. Thus the problem is that personal equipoise may be lost more readily than societal equipoise.

If as an individual we can make a rational decision to go with treatment A, how can we as investigators then turn around and ask a patient to submit to a randomization device that may well decide that that patient gets treatment B? This distinction came to the fore in the classical case of the acceptance of the use of extracorporeal membrane oxygenation (ECMO) to treat newborns with persistent pulmonary hypertension in the UK. Possibly influenced by the costs of the procedure, the authorities were not convinced by the experience in the USA and demanded, and mounted, a randomized clinical trial to settle the issue. The evidence from the USA included trials from Michigan[10] where 100 babies were studied, and a randomized study[11] where one control baby died and eleven babies on ECMO survived. There was a two phased randomized study from Children's Hospital Boston[12] where in the first phase six of ten control babies survived and nine of the nine babies on ECMO survived.

[10]R Bartlett et al.: Extracorporeal membrane oxygenation (ecmo) in neonatal respiratory failure. 100 cases. In: Annals Surgery 204 (1986), pp. 236–45.

[11]R Bartlett et al.: Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study, in: Pediatrics 76 (1985), pp. 479–87.

[12]P O'Rourke et al.: Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the newborn: a prospective randomized, in: Pediatrics 84 (1985), pp. 957–63.

In the second phase twenty babies were treated with ECMO and nineteen survived. These trials were supported by a register of experience with 715 babies on ECMO between the years 1980-1987 in 18 neonatal centers in the USA[13]. To their credit the UK study was stopped early, but 54 of the 92 babies placed on the control arm died (in contrast to 30 of the 93 on ECMO). It is difficult to believe that any informed, caring parent of a child born prematurely and eligible for this study would not choose ECMO for their child.

The argument against revealing the accumulated knowledge of the study as it progresses is that this action would compromise the study's scientific basis. With studies where psychological considerations, such as the placebo effect, for example, are important and blinding is used, such a disclosure might well invalidate subsequent data. For example, an early clinical trial, designed by Benjamin Franklin and Antoine Lavoisier to expose Anton Mesmer[14], found it necessary to actually have the patients blindfolded; and thus the origin of the use of the word blind in the description of studies. They also lied to the patients, tricking them into believing they were receiving certain treatments that they were not, but that should not happen today because of informed consent. Informed consent is what makes modern trials ethical. It is argued in[15] that informed consent is certainly necessary, but is not sufficient for an ethical trial, and amongst other conditions necessary to make a trial ethical is respect for the patient. This respect includes keeping the informed consent document current, which means including information that may only become available during the performance of the trial. Surely the patients should be kept informed of information that becomes available in an interim analysis that may impact on patients' judgment of equipoise, otherwise the informed consent form is no longer informed. This may cause problems for those analyzing the data from such trials, but only if a patient armed with the new information still wishes

[13] J Toomasian et al.: National experience with extracorporeal membrane oxygenation for newborn respiratory failure. data from 715 cases, in: ASAIO-transactions-American-Society-for-Artificial-Internal-Organs 34 (1988), pp. 140–7.

[14] H Herr: Franklin, lavoisier, and mesmer: origin of the controlled clinical trial, in: Urologic Oncology 23 (2005), pp. 346–351.

[15] E Emanuel/D Wendler/C Grady: What makes clinical research ethical?, in: Journal of the American Medical Association 283 (2000), pp. 2701–2711.

to be randomized, is it an informed decision. Furthermore, with our statistical techniques, we demonstrate that it is possible to obtain unbiased estimates for situations where patients may choose randomization after a non-significant interim analysis has occurred.

2.3 Materials and methods

2.3.1 Scenario

We identified the following case as a stereotypical problem that highlights the issues at the heart of the general problem, and then proceeded to propose unbiased estimators that were valid even in this irregular situation.

Consider the following example to typify the situation: there are two arms in the trial, one and two, with the intended treatments of A and B, respectively, with 400 people in each arm. The trial is designed to allow one interim analysis at 50% information (300 deaths) and one final look at 100% information (600 deaths). We model patient accumulation to be such that every 30 days 100 more people (50 in each arm) are added to the trial until the trial has 800 patients. Our aim is to estimate the hazard ratio of treatment B versus treatment A.

At the beginning of the trial we had some form of equipoise, so it was ethical to randomly assign patients to each treatment. At the interim-look, if the analysis showed that the difference between the treatments was significant then we would terminate the trial, whereas if the difference was not deemed significant, then we wouldl continue with the trial. However, we may no longer have equipoise as one of the treatments will have been observed as possibly better, even though the degree to which it is better is not convincing to the societal norms that define a "significant" difference. If this were the case, then suppose this new information gained at the interim analysis were to be incorporated into the informed consent form giving the patients subsequently randomized to the "inferior" (but not significantly so) arm the option to choose the other treatment.

To be more precise, we define the triplet $A/B/C$ to denote the hazard where:

- A is either G or W, denoting if the patient is randomised to the better (good) or worse treatment
- B is either $D.Better$ or $D.Current$, denoting the desire to switch to the apparently better treatment if offered, or to stay on whatever treatment they were randomized to
- C is either $H.Stay$ or $H.Switch$, denoting whether or not the person has switched treatments

The main rationale for randomization is to maintain balance between treatment arms; allowing patients to choose other than the randomly assigned treatment may introduce biases, which may not be measurable. To quantify the situation, before the interim look, we had four hazards operating in the trial: those who were given treatment A and would remain on it regardless ($G/D.Current/H.Stay$), those who were given treatment A and would switch given the new information (not yet available) from the interim analysis ($G/D.Better/H.Stay$), those who were given treatment B and would remain on it ($W/D.Current/H.Stay$), and those who were given treatment B and would switch ($W/D.Better/H.Stay$). After the interim look, we had three hazards operating for any new patients in the trial: those who were given treatment A and would remain on it ($G/D.Current/H.Stay$), those who were given treatment B and would remain on it ($W/D.Current/H.Stay$), and those who would switch if given the chance (both $G/D.Better/H.Stay$ and $W/D.Better/H.Switch$, which were the same hazards). We did not think it reasonable to switch from the assigned treatment to an apparently "lesser" treatment, so we ignored this possibility.

Our problem then involved estimating the true hazard ratio between the two treatments in the above scenario where, given a non-significant interim look and proceeding with the trial, those on the supposed inferior arm were given the option to switch to the superior arm.

2.3.2 Basic principles

The idea of asking patients which treatment they would prefer has previously been broached by Zelen in[16]. Zelen argues that allowing patients to choose their treatment, and subsequently performing standard analyses, would possibly result in a lack of statistical efficiency, however, estimates would ultimately be mostly unbiased if few patients opt to switch treatments. This, however, does not address the situation where large quantities of patients decide to switch treatments - a case that our method addressed.

Our method of analysis relied on a Bayesian argument using Monte Carlo methods: we hypothesised realistic values for the aforementioned four hazards, and simulated a large number of trials for which we know the true hazard ratios. In doing so, we observed the distribution of estimated hazard ratios at the final look, given the true hazard ratio; that is, P(estimated final HR|true HR).

That is, we took a wide selection of possible values for the hazards $G/D.Current/H.Stay$, $G/D.Better/H.Stay$, $W/D.Current/H.Stay$, and $W/D.Better/H.Stay$. In future applications, this could be done at the interim look by finding the mean time to death for each arm (for similar trials), and inverting it to find the rates. Reasonable maximum and minimum bounds of the hazards could be obtained by four times the maximum estimated rate, and one-quarter of the minimum estimated rate, respectively. Values could then be selected at equal spacing between the aforementioned bounds.

For each selection, we simulated the aforementioned study plan: one interim analysis at 50% information (300 deaths) and one final look at 100% information (600 deaths). We modelled patient accumulation to be such that every 30 days 100 more people (50 in each arm) were added to the trial until the trial had 800 patients. The option of new patients choosing their treatment was possible after the interim analysis.

For each selection of hazards, an estimated final hazard ratio was produced by analyzing

[16]M Zelen: A new design for randomized clinical trials, in: The New England Journal of Medicine 300 (1979), pp. 1242–1245.

the study data after 600 deaths in the normal fashion. This estimate was obviously biased. We also generated the true unbiased hazard ratio by simulating the same study, but without any options of treatment choice. We then established a database of relationships (for each set of hazards $H.SET = [G/D.Current/H.Stay, G/D.Better/H.Stay, W/D.Current/H.Stay,$ and $W/D.Better/H.Stay]$) between the true unbiased hazard ratio and biased estimates that we obtained through allowing switching to occur. This allowed us to see the probabilistic relationship between the biased estimates and the true estimates; that is, we created an estimation of P(estimated final HR|true HR). It will aid the reader to imagine the database as a rectangular matrix with the true hazard ratio values as columns and the estimated final hazard ratio values as entries in the applicable column.

In a simplied case, we had the hazard sets:

$$H.SET_a = [a_1, a_2, a_3, a_4]$$
$$H.SET_b = [b_1, b_2, b_3, b_4]$$

Which produced the true hazard ratios (when no switching is allowed):

$$HR(H.SET_a)$$
$$HR(H.SET_b)$$

We then simulated the hazard ratios that occurred when switching was allowed (here for the sake of simplicity, two simulations per set are displayed, but in practice this number was 10,000):

$$[HR_a(1), HR_a(2)] \in H.SET_a$$
$$[HR_b(1), HR_b(2)] \in H.SET_b$$

We then organized this in the shape of the aforementioned rectangular matrix, as shown in

Table 2.1.

Table 2.1. Hypothetical database of hazard ratio relationships

$HR(H.SET_a)$	$HR(H.SET_b)$
$HR_a(1)$	$HR_b(1)$
$HR_a(2)$	$HR_b(2)$

A hypothetical database of hazard ratios, where true sets of hazards are stored in the first row, and the observed (biased) hazard ratios are stored in the rows beneath

We estimated $P(\text{true HR}|\text{estimated final HR})$ by restricting the simulated dataset to those with estimated final hazard ratios within a certain range of interest (i.e. conditioning on the estimated final HR), and formed a probability density function of the counts of the true hazard ratios corresponding to those observations left remaining[17].

2.3.3 Naive approach

Problems did arise, as results were only stored if the interim analysis is non-significant, and non-significant results were a lot more likely to occur when the true hazard ratio is closer to one. Thus if we only calculated the posterior distribution by observing the count of similar estimates in each true hazard ratio category and then normalizing the distribution to sum to one, the estimate would be biased towards the null. Because of this, we used the proportion of observations in each true hazard ratio category as a measure of an estimate's frequency in a particular true hazard ratio category, and then normalized all of the proportions to sum to one.

To restate this problem mathematically, the probability that a trial (with a hazard ratio HR) would not be terminated early, p_{HR}, (and thus eligible for inclusion in the stored matrix) was a monotone function that had a positive relationship to the true hazard ratio (only the

[17]S Jackman: Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo, in: American Journal of Political Science 44 (2000), pp. 375–404.

region where the hazard ratio is less than one is considered):

$$p_{HR} = f(HR); \text{where } f(x) < f(y) \text{ if } x < y.$$

Consider the simple case where we only simulate two possible true hazard ratios: 0.8 and 0.4. We know that

$$p_{0.6} = f(0.6) < f(0.8) = p_{0.8}$$

so if we simulate 10,000 trials for each hazard ratio, we can expect $10,000 \times p_{0.6} = 500$ and $10,000 \times p_{0.8} = 8,000$ results in the matrix for true hazard ratios of 0.6 and 0.8 respectively. Thus if we have a true hazard ratio of 0.69 (the exponent geometric mean of 0.6 and 0.8) for an unbiased estimate we require equal representation in the matrix columns of 0.4 and 0.8. As it is equally likely that the estimate for 0.69 will show up in each column of 0.6 and 0.8, there will be 16 times more observations in the 0.8 column than in the 0.6 column. This means with regards to estimates:

$$P(0.8|\text{estimate of } 0.69) = 16 \times P(0.6|\text{estimate of } 0.69)$$

and thus

$$E[\text{True HR}|\text{estimate of } 0.69] = \exp(16 \times \log(0.8) + 1 \times \log(0.6))/17) = 0.79$$

It is clear that the estimate is strongly biased towards the less-extreme observation.

If we instead use the prevalence of the observation in each column, then

$$P(0.8|\text{estimate of } 0.69) = P(0.6|\text{estimate of } 0.69)$$

and thus

$$E[\text{True HR}|\text{estimate of } 0.69] = \exp(\log(0.8) + \log(0.6))/2) = 0.69$$

resulting in an unbiased estimate.

Problems further arose when we noted that we were not performing a simple comparison of two hazards, but instead four. When simply comparing two hazards, it is a simple task to hold one constant (e.g. $hazard_1 = 0.01$) and vary the other (e.g. $hazard_2 = 0.01, 0.012, 0.014, \ldots, 2$), resulting in a uniform spread of hazard ratios to sample from, while maintaining a fair sampling method with regards to the hazards. When comparing four hazards ($G/D.Current/H.Stay$, $G/D.Better/H.Stay$, $W/D.Current/H.Stay$, and $W/D.Better/H.Stay$), however, ensuring both a fair sampling method of the hazards and a uniform spread of resultant hazard ratios to sample from is much harder.

We circumvented this problem by implementing an even uniform spread of the four hazards over a predetermined range (e.g. $G/D.Current/H.Stay$, $G/D.Better/H.Stay$, $W/D.Current/H.Stay$, and $W/D.Better/H.Stay \in$ [0.01, 0.012, 0.014, ..., 0.02]), calculated the true hazard ratio for each permutation of hazards (HR), assigned these hazard ratios into a predetermined uniform spread of hazard ratio bins (e.g. $[0 - 0.05], (0.05 - 0.1], (0.1 - 0.15], \ldots, (0.95 - 1]$), and performed all calculations using the spread of bins.

As hazard ratios have all the properties of ratios, it was preferable to create this aforementioned predetermined uniform spread of hazard ratio bins as being uniformly spaced in the log scale. The sampling was also performed on the log scale so the sampling distribution was less likely to be skewed, and thus the mean was more appropriate.

The mean was also preferable to the median because the median would only return values already in the distribution. If the true hazard ratio was not specified in the hazard ratio bins, no amount of statistical trickery would result in an unbiased estimate. That is, if we had two bins that were equally distributed on the log scale: [0.2-0.45] and (0.45-1], with

corresponding midpoints of 0.30 and 0.67, then to estimate the true hazard ratio we would sample from each of these bins 10,000 times. If we used the median as the desired estimation measure, then our unbiased estimate could only be 0.30 or 0.67, whereas if we used the mean, then any range of estimates were possible.

Performing this naive method resulted in less biased observations than if we analyzed the data blindly, and the more simulations initially ran and stored in the database of P(estimated final HR|true HR) the better the observations. Taking another logical step, by adding even more information to the estimation process, it was possible to get even better estimates.

2.3.4 Additional information and priors

Information was added by conditioning on more information than just the estimated final hazard ratio; we conditioned upon the estimated interim hazard ratio. That, is:

$$P(\text{true HR}|\text{estimated interim HR \& estimated final HR}).$$

Conditioning on this additional hazard ratio – as opposed to just those previously suggested – added more information to the sampling process, and allowed us to better estimate the true hazard ratio[18].

When we noticed trends in the bias (e.g. the estimates may have been biased towards the null in situations where the true hazard ratio was extreme, and biased away from the null when the true hazard ratio was close to the null) we added a prior to the Bayesian sampling method[19]. Priors are generally useful for overarching adjustments to the model, as opposed to trying to account for additional variation in the model. As our interim look was the least biased estimate we had (it is not completely unbiased however, as it was biased towards

[18]Jackman: Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo (see n. 17).
[19]Ibid.

the null because continuing with the trial was conditional upon the interim look being non-significant) we implemented our prior around the interim estimation. In this example, when the interim estimate was closer to the null, we sampled more from the less-extreme tail, and when the interim estimate was further away from the null (say a hazard ratio of 0.85 or less) then we sampled more from the more-extreme tail.

The downside to these additional processes was that we required much more information in the simulation database, which was costly both in storage and time. The time cost was not only expensive in terms of simulating the initial database, but also when trying to optimize the priors (a process that might take many tries) each simulation took much longer as it is more computationally intensive. We sought to find the optimal balance between the amount of information in the model and the amount of time we had[20]. Of course, computer time is trivial when compared to the human cost of clinical trials.

2.3.5 Hypothesis testing

It is also important to realize that the user can still implement hypothesis testing when using this bias-adjusted technique. By sampling from the posterior distribution, $P(\text{true HR}|\text{information})$, it is possible to get confidence intervals for the true hazard ratio for any level of confidence, thus allowing hypothesis testing to take place[21]. If one desires p-values as opposed to seeing if their confidence interval contains the null, then it is possible to derive the p-value by determining at what quantile the null lies, and then subtracting the value from one, and halving the resulting value (if a two sided test is desired, otherwise the halving is not required).

[20] J Stewart: Optimization of parameters for semiempiral methods i. method, in: Journal of Computational Chemistry 10 (1989), pp. 209–220.

[21] T DiCicco/B Efron: Bootstrap confidence intervals, in: Statistical Science 11 (1996), pp. 189–228.

2.4 Results

2.4.1 Naive application

We modelled the previously mentioned scenario, with 50% of the new patients switching treatments if given the chance. We started by generating our own database: 10,000 trials for each combination of hazards, with one hazard rate (arm one, those who do not switch) fixed at 0.01/day, another (arm two, those who do not switch) varied in 0.001 increments from 0.005 to 0.027, and the final two hazard rates (those who switch in both arms) varied from -0.004, -0.002, 0, 0.002, and 0.004 in an additive manner from the hazard rates of those who stayed, in their arms of the trial, respectively. The estimation of these 575 hazard ratios (both the interim and final estimates were saved) took approximately a week running on a 3.2 GHz 64 bit core with 6 Gb of RAM.

We then proceeded to (again) simulate a range of hypothetical trials, in order to test our method's ability to correctly estimate them. We generated 1,000 trials for each combination of hazards, with one hazard rate (arm one, those who do not switch) fixed at 0.01/day, another (arm two, those who do not switch) taking the values 0.01, 0.011, 0.0125, 0.01428, 0.01667, and 0.02 (representing hazard ratios of 1, 0.9, 0.8, 0.7, 0.6, and 0.5 respectively). The final two hazard rates (those who switch in both arms) varied through -0.002, 0, and -0.002 in an additive manner from the hazard rates of those who stayed in their arms of the trial, respectively.

One of the issues here was the width of the area around an estimate for it to result in it being observed within a particular bin, as defined earlier in the paper. After trial and error, it was decided that the bins should be equally spaced for both the interim and final looks. Common sense would have had us place smaller widths around the interim estimate, as it had everyone abiding by their correct treatment. However, there were also less people present, and hence less information than in the final estimate, so equal spacing was most appropriate. Of course, after observing the fit of the model with equal spacing, it is then

the investigator's prerogative to alter the width as appropriate.

On applying the naive Bayesian method, modelling

$$P(\text{true HR}|\text{estimated interim HR \& estimated final HR})$$

we obtained the results present in Figure 2.1.

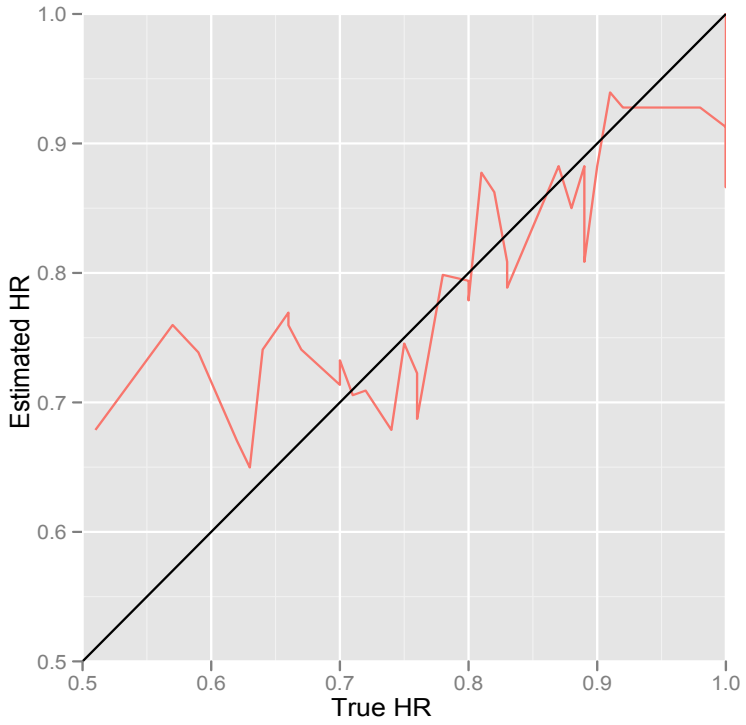

Figure 2.1. Results from the naive Bayesian method

2.4.2 Application with additional information

The main problem with Figure 2.1 is the lack of information; we saw that there was some crucial element of the data that we were not capturing, as there was an underlying variable that caused the estimated hazard ratio to bounce up and down in a non-systematic manner. We concluded that our current model, only conditioning upon the interim and final estimates of the hazard ratio, was not capturing the subtleties of how those who would switch

treatments have different hazard ratios from those who would stay. We needed to account for this, yet we had to be mindful of the computational burden of conditioning on extra variables, so we generated (for each arm of the trial at the interim look) the hazard ratio of those who would switch, compared to those who would stay.

We addressed this problem of lack of information by generating the aforementioned hazard ratios, and by conditioning on these hazard ratios we modelled

$$P(\text{true HR}|\text{estimated interim HR \& estimated final HR \&}$$
$$\text{estimated interim HR restricted to only those who would switch \&}$$
$$\text{estimated interim HR restricted to only those who would not switch})$$

and achieved the fit present in Figure 2.2.

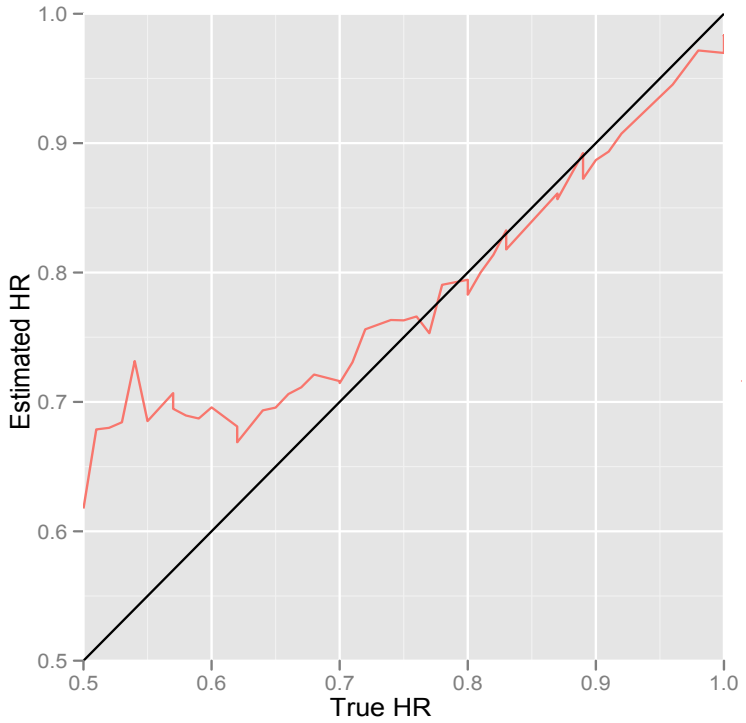

Figure 2.2. Results from the naive Bayesian method with additional information

It was readily apparent that there was much less randomness to the new model, as

we accounted for the variability by conditioning upon the hazard ratios between those who switched and those who stayed, within each arm. That is, by including additional information into the model, we gained a smoother (and better) fit.

2.4.3 Application with priors

The astute observer can note that there appears to be a trend where the estimates are biased towards the null when the true hazard ratio is more extreme, and biased away from the null when the true hazard is less extreme. This is because we discarded results that were significant at the interim look (which is much more likely to occur in the true extreme cases) and then normalized the results to sum to one. This was fixed by adding a prior to the Bayesian sampling method[22]; when the interim estimate was closer to the null, we sampled more from the less-extreme tail, and when the interim estimate was further away from the null (a hazard ratio of 0.85 or less) we sampled more from the more-extreme tail.

After applying priors to make the observations less extreme for interim observations less extreme than 0.84, and more extreme for interim observations between 0.70 and 0.75, and even more extreme for interim observations below 0.70 results in the fit present in Figure 2.3.

2.4.4 Bias

By applying our method, we obtained functionally unbiased estimates of the true hazard ratio. When the true hazard ratios varied between 1 and 0.55 (most practical clinical scenarios fall within these bounds), we observed that our bias was within ± 0.03, as shown in Figure 2.4.

Within 0.65 to 1, we controlled the majority of the bias to be within ± 0.01, and for all intents and purposes all of the bias was within ± 0.02, as shown in Figure 2.5. However, it is

[22] Jackman: Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo (see n. 17).

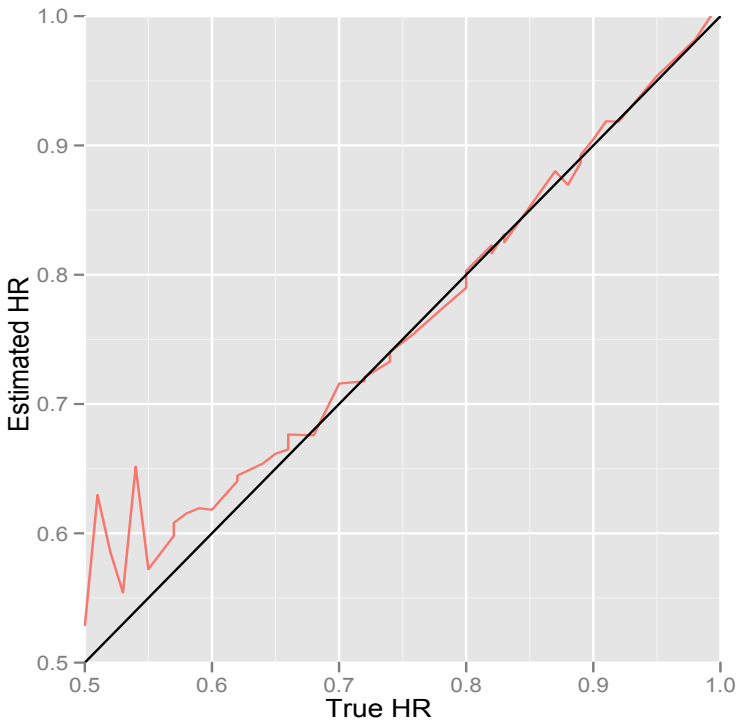

Figure 2.3. Results from the Bayesian method with priors applied

worth considering that in a real clinical trial, if the hazard ratio were that extreme, then it is most likely that the interim analysis would have been significant and the trial terminated, removing the need for our method.

Between 0.55 and 0.5, our estimates started to waiver due to our comparable lack of information in these areas; given a larger simulation database, these results would be less biased; similar to those between 1 and 0.55.

2.5 Discussion

By using our method it was shown to be possible to obtain functionally unbiased results, allowing us to statistically rescue a biased scenario in which a traditional analysis was "sacrificed" in order to treat patients ethically.

Let us now consider historical clinical trials and start by taking the preventative breast

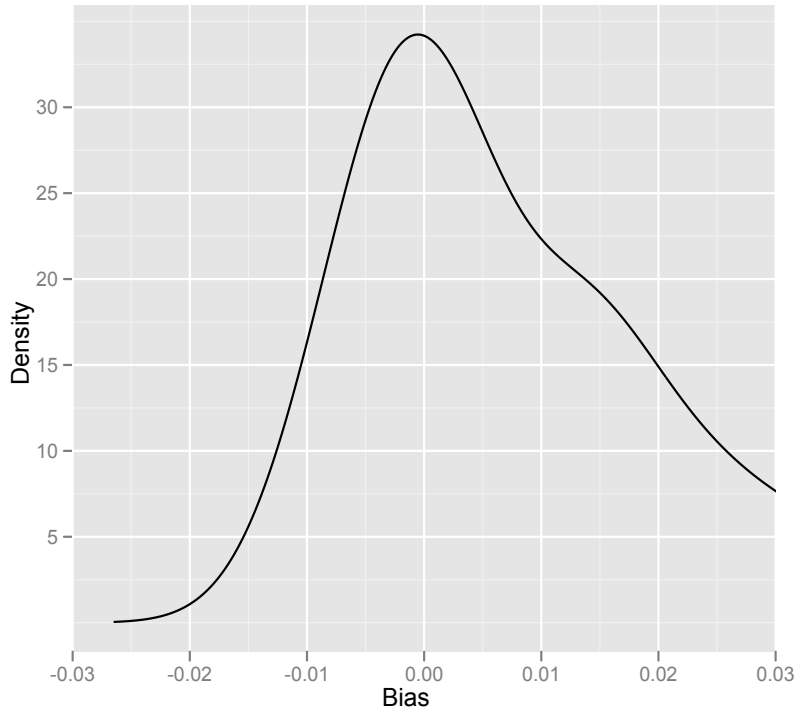

Figure 2.4. Distribution of the bias when the true hazard ratio lies between [0.55,1]

cancer trial for Exemestane[23]. By the end of the trial, 11 patients on Exemestane had developed invasive breast cancer, as opposed to 32 on placebo. If the trial conductors had observed that Exemestane had been slightly more effective at a 50% interim look, which we can assume happened halfway through trial recruitment, and then half of the remaining patients to be recruited to placebo had decided to switch to Exemestane, an estimated six cases (eight less cases in the placebo arm, two more added in the Exemestane arm) of breast cancer could have been averted. That is, the total number of cases of breast cancer in this trial could have been reduced from 43 to 37; a reduction of 14% of women in this trial from developing breast cancer, with no additional cost or logistical effort to the study.

Similarly, in this trial for preexposure chemoprophylaxis for HIV prevention[24] by the end

[23]P Goss et al.: Exemestane for breast-cancer prevention in postmenopausal women, in: The New England Journal of Medicine 364 (2011), pp. 2381–2391.

[24]R Grand et al.: Preexposure chemoprophylaxis for hiv prevention in men who have sex with men, in: The New England Journal of Medicine 364 (2010), pp. 2587–2599.

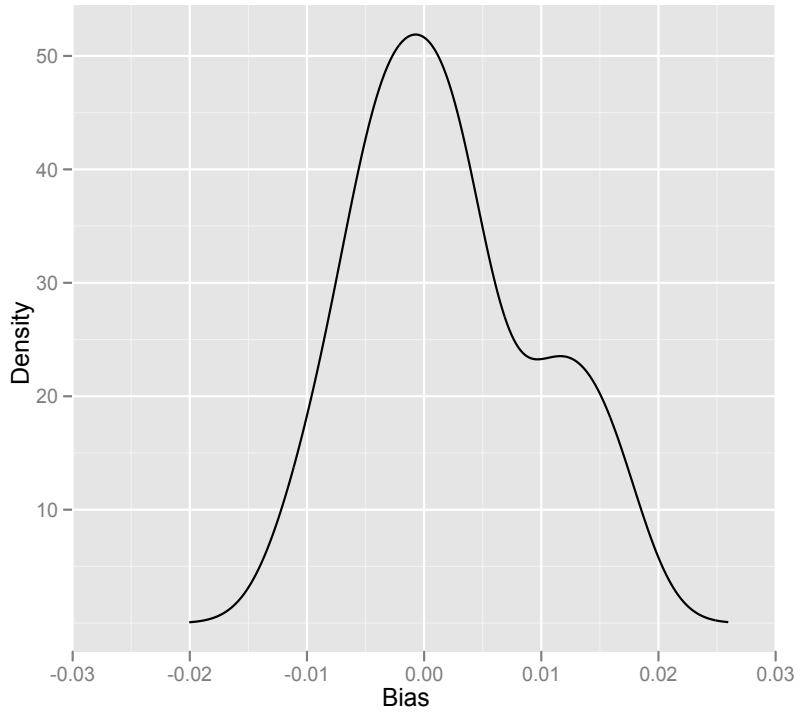

Figure 2.5. Distribution of the bias when the true hazard ratio lies between [0.65,1]

of the trial 36 patients on antiretroviral chemoprophylaxis became infected with HIV, and 64 on placebo were likewise infected. With the same assumptions as in the Exemestane trial, ten people (sixteen less cases from the placebo arm, six more added to the antiretroviral arm) could have avoided being infected with HIV. Again, decreasing the number of infected from 100 to 90, a reduction of 10% of men in this trial from being infected with HIV, with no additional cost or logistical effort to the study.

True informed consent requires transparency; that is, current information. If the trial has not been terminated early, then information from the interim analysis must be used to update the informed consent. It is possible to achieve this by implementing our method, and doing so has the potential to stop many needless adverse events in the modern clinical trial setting. By altering the statistical analysis process it was possible to save six women from breast cancer, and ten men from HIV infection, in just two clinical trials. These numbers become even more important when multiplied across the vast number of clinical

trials performed each year.

To the best of our knowledge, this is the first paper to propose and validate a random clinical trial protocol that would allow subjects to choose their own treatments, while still producing unbiased hazard ratio estimates that can be used in hypothesis tests.

Chapter 3

Novel developmental analyses identify longitudinal patterns of early gut microbiota that affect infant growth

3.1 Abstract

It is widely acknowledged that obesity trajectories are set early in life, and that rapid weight gain in infancy is a risk factor for later development of obesity. Identifying modifiable factors associated with early rapid weight gain is a prerequisite for curtailing the growing worldwide obesity epidemic. Recently, much attention has been given to findings indicating that gut microbiota may play a role in obesity development. This is the first longitudinal study that aims at identifying how the development of early gut microbiota is associated with expected infant growth. We developed a novel procedure that allows for the identification of longitudinal gut microbiota patterns (corresponding to the gut ecosystem evolving), which are associated with an outcome of interest, while appropriately controlling for the false discovery rate. Our method identified developmental pathways of Staphylococcus species and Escherichia coli that were associated with expected growth. Our method should have wide

future applicability for studying gut microbiota, and is particularly important for translational considerations, as it is critical to understand the correct timing and design of the interventions, prior to attempting to manipulate gut microbiota in early life.

3.2 Introduction

Gut microbiota has a critical role in human health[1]; early infancy is of special interest because the early life period is a determinant for the subsequent adult-like microbiota. Once the first microbes arrive in the sterile gut of the newborn, a dynamic process starts, where activation of genes and expression of receptors in the host plays an important role for the building of niches and the further selection of microbes. More importantly, studies on germ free animals have revealed the presence of time-dependent exposure windows that rely on microbial stimuli from the gut[2] (i.e. development of tolerance[3], sensitivity to biogenic amines[4], influences on cecum size[5], and optimal functioning of diverse systems, such as angiogenesis[6] and stress responses[7]).

Obesity has been linked to gut microbiota in humans, by being associated with reduced

[1]F Backhed et al.: Host-bacterial mutualism in the human intestine, in: Science 307 (2005), pp. 1915–20; T Mitsuoka: Intestinal flora and aging, in: Nutrition Reviews 50 (1992), pp. 438–46; J Roung/S Mazmanian: The gut microbiota shapes intestinal immune responses during health and disease. In: Nature Reviews Immunology 9 (2009), pp. 313–23; PJ Turnbaugh et al.: A core gut microbiome in obese and lean twins, in: Nature 457 (2009), pp. 480–4; B Bjorksten et al.: The intestinal microflora in allergic estonian and swedish 2-year-old children, in: Clinical and Experimental Allergy 29 (1999), pp. 342–6; S Mazmaniam/J Round/ D Kasper: A microbial symbiosis factor prevents intestinal inflammatory disease. In: Nature 453 (2008), pp. 620–5.

[2]C Thompson/B Wang/A Holdes: The immediate environment during postnatal development has long-term impact on gut community structure in pigs, in: ISME 2 (2008), pp. 739–48.

[3]N Sudo et al.: The requirement of intestinal bacterial flora for the development of an ige production system fully susceptible to oral tolerance induction, in: The Journal of Immunology 159 (1997), pp. 1739–45; S Mazmanian et al.: An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system, in: Cell 122 (2005), pp. 107–18.

[4]B Gustafsson/T Midtvedt/K Strandberg: Effect of microbial contamination on the cecum enlargement of germ free rats, in: Scandinavian Journal of Gastroenterology 5 (1970), pp. 309–314.

[5]Ibid.

[6]T Stappenbeck/L Hooper/J Gordon: Developmental regulation of intestinal angiogenesis by indigenous microbes via paneth cells, in: Proceedings of the National Academy of Sciences of the United States of America 99 (2002), pp. 15451–5.

[7]N Sudo et al.: Postnatal microbial colonization programs the hypothalamic-pituitary-adrenal system for stress response in mice, in: The Journal of Physiology 558 (2004), pp. 263–75.

bacterial diversity and altered representation of bacterial genes and metabolic pathways[8]. Since rapid weight gain in early life is a risk factor for the later development of obesity[9], we aimed to study whether early infant gut microbiota was associated with expected growth in the first six months of life. As gut microbiota can be altered, or even transplanted[10], there is large potential for future medical interventions.

We describe a novel method that identifies patterns of gut microbiota exposures associated with potential time-dependent exposure windows in longitudinal data. We implement this method in the Norwegian Microflora Study (NOMIC) to reveal which patterns of gut microbiota (representing the gut ecosystem evolving) are associated with expected infant growth.

This is the first longitudinal study that aims at identifying how the development of early gut microbiota affects infant growth. Proper knowledge of the time dependencies of gut microbiota as an exposure is a crucial underpinning before experimental attempts to manipulate early gut microbiota can be made. In light of this, our method will have considerable future applications, especially in the translational area of gut microbiota research.

3.3 Materials and methods

3.3.1 Study population

NOMIC is a birth cohort designed to study the establishment of gut microbiota during infancy and its consequences for child health. Participating mothers were recruited to the NOMIC study by a paediatrician at the maternity ward in a county hospital in South Norway. The recruitment protocol purposefully oversampled preterm children; whenever a preterm-birth mother was enrolled, two mothers of consecutively born term infants were recruited.

[8]Turnbaugh et al.: A core gut microbiome in obese and lean twins (see n. 1).

[9]P Monteiro et al.: Birth size, early childhood growth and adolescent obesity in a brazilian birth cohort, in: International Journal of Obesity 27 (2003), pp. 1274–82.

[10]Turnbaugh et al.: A core gut microbiome in obese and lean twins (see n. 1).

The recruitment started in November 2002 and was completed in May 2005. Eligibility criteria required that mothers were fluent in Norwegian and a resident in the pertinent geographic area. The study was approved by the Norwegian Data Inspectorate and the Regional Ethics Committee for Medical Research.

After the informed consent forms were signed by the mothers, containers for fecal samples and a questionnaire were provided to the participants at the maternity ward. The mothers were asked to collect and freeze one fecal sample from themselves at postpartum day 4, as well as samples from their infants when they were 4, 10, 30, and 120 days old. Study personnel retrieved the fecal samples and kept them frozen during transport to the Biobank of the Norwegian Institute of Public Health, Oslo, where they were stored at -20 C upon arrival. Further questionnaires were sent to the families when their infants were aged 6, 12, 18, and 24 months.

Six hundred and one mothers agreed to participate in the NOMIC study, however, 86 (14%) of these mothers never returned any fecal samples, which left 524 infants with available fecal samples from one or more occasions. Children that were preterm (152), born via caesarean section (169), or had been exposed to antibiotics before day 4 of life (124), were then excluded from the current analysis, leaving 246 children.

3.3.2 Outcome

Mothers extracted information on weight from their "baby health visit" cards and reported this information in questionnaires. Information on gestational age and preterm delivery was obtained from the Medical Birth Registry of Norway.

To be included in the analysis, we required birthweight and another weight measurement within 122 to 244 days of birth (approximately 4 to 8 months). These two measurements are henceforth referred to as measurements at birth and six months of life. If multiple measurements were available during the latter period, the closest to 6 months was used. Data from 218 children (110 females and 108 males) met the inclusion criteria.

The infants' weights were expressed as an age and sex standardised Z-score. Following recommendations from the Norwegian Health Directorate[11], we used the World Health Organisation's weight-for-age growth curves[12] to generate these Z-scores. The Z-score of the weight at birth (Z_{0i}) was compared to the Z-score of the child at six months of life (Z_{6i}). We defined the outcome of interest to be the difference in Z-scores: $Y_i = Z_{6i} - Z_{0i}$. This definition was chosen to be in concordance with the current literature, where the most frequent definition of rapid growth was a Z-score change in weight-for-age[13]. If a child's Z-score deviated between time periods, it was indicative of deviant growth and labelled as either increased growth (reaching higher weights than expected from its birthweight) or decreased growth (undershooting the target weight and reaching lower weights than expected).

The distribution of the difference in Z-scores was found to be approximately Normally distributed, with a mean of -0.29, median of -0.38, and IQR of -0.80 to 0.24 for females, and a mean of -0.13, median of -0.18, and IQR of -0.82 to 0.57 for males. To aid in the interpretation of Z-scores, the relationship (at different birth weights) between change in Z-score and weight at six months is displayed in Figure 3.1. Table 3.1 contains further descriptive characteristics of the study participants.

3.3.3 Exposures

16S rRNA gene clone libraries were constructed from DNA extracted from the fecal samples obtained on days 4, 10, 30, and 120. Detailed information about this process can be found in a previous paper from the NOMIC study[14].

The exposures of interest are intensity readings for 22 probes, encoding different gut

[11] Arbeidsgruppe: Nasjonale faglige retningslinjer for veiing og maaling i helsestasjons - og skolehelsetjenesten, Oslo, Norway 2010.

[12] World Health Organization: WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development, Geneva, Switzerland 2006.

[13] P Monteiro/C Victoria: Rapid growth in infancy and childhood and obesity in later life - a systematic review, in: Obesity 6 (2005), pp. 143–154.

[14] M Eggesbo et al.: Development of gut microbiota in infants not exposed to medical interventions, in: Acta Pathologica, Microbiologica et Immunologica Scandinavica 119 (2011), pp. 17–35.

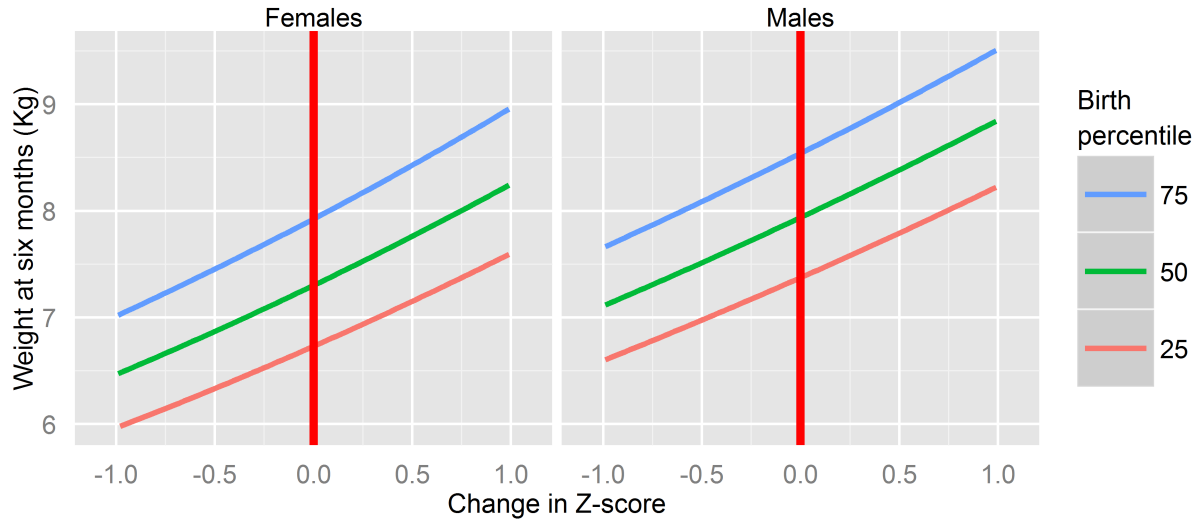

Figure 3.1. Relationship between weight at six months and change in Z-score. The relationship is displayed for multiple birthweight percentiles. A change in Z-score of 0 corresponds to theoretically perfect growth at six months. If a male child was born at the 75th percentile, then their expected weight at six months would be 8.5 kg (y-axis), corresponding to 0 change in Z-score (x-axis) on the right panel. If the child instead weighed 9.0 kg at six months (y-axis), then that would correspond to a +0.5 change in Z-score (x-axis).

microbiota species (spp.) groups at 4, 10, 30, and 120 days since birth. The probes, labelling sequence, and target bacteria spp. groups are displayed in Table 4.2. The frequency of each probes detection, stratified by day and sex, are shown in the appendix (Figure A.1).

Each intensity reading at every time point is dichotomised into either detected or non-detected. We selected this categorisation since we had no information on the distributions of the different probes' intensities in the average population, i.e. it was not possible to choose appropriate demarcations for low, moderate, or high levels.

Each microbiota spp. group were examined individually.

3.3.4 Confounders and effect modification

Information on potential confounders was obtained by questionnaires filled in by the mothers and from the Medical Birth Registry of Norway. Variables considered *a priori* to be potential

Table 3.1. Descriptive characteristics of study participants

Characteristic	Description
Maternal smokers (%)	11.5
Twins (%)	3.3
Siblings (%)	61.8
Birthweight (Kg)	3.58 (3.27, 3.88)
Gestational age (days)	284 (277, 288)
Maternal age (yrs)	30 (28, 33)
Maternal BMI	24 (21, 26)
Sample size	218

Statistics are displayed as median (IQR) or only %. Sex specific results were not noticeably different from the above results.

confounders were antibiotics use (after day 4 of life), sex, milk substitutes, maternal smoking, and parity, however, stepwise regression procedures led to the removal of all considered confounders due to a lack of effect.

When considering the relationship between microbes and growth, our initial investigations found evidence for effect modification by sex. This led us to perform separate stratified analyses.

3.3.5 Time-specific analyses

We were interested in identifying time points at which the detection of specific gut microbiota spp. groups were significantly associated with growth trajectory. That is, we investigated whether we could identify any time points, where the detection of gut microbiota spp. groups, shifted the growth outcome, the mean change in Z-score. We modelled this relationship by including the detection of gut microbiota at each time point (days 4, 10, 30, and 120) separately, using a standard linear regression model (separately for every gut microbiota spp. group):

$$\begin{aligned} Y_i =& \beta_{0,j_k} + \beta_{1_k} \cdot X_{i,4_k} + \beta_{2_k} \cdot X_{i,10_k} + \beta_{3_k} \cdot X_{i,30_k} + \\ & \beta_{4_k} \cdot X_{i,120_k} + \epsilon_{i,j_k}, \end{aligned}$$

Table 3.2. Probes and their targets

#	Probe match	Labelling
1	*Enterococcus* spp.	TCATCCCTTGACGGTATCTAA
2	*Lactobacillus* spp.	GTCAAATAAAGGCCAGTTACTA
3	*Lactobacillusi paracasei/case*	CAGTTACTCTGCCGACCATT
4	*Staphylococcus* spp.	ACACATATGTTCTTCCCTAATAA
5	*Streptococcus* spp. *(α-hemolytic)*	AGTGTGAGAGTGGAAAGTTCA
6	*Clostridium* spp.	TCAACTTGGGTGCTGCATTC
7	*Lachnospiraceae* spp.	AGCTAGAGTGTCGGAGAGG
8	*Veillonella* spp.	GATTGGCAGTTTCCATCCCAT
9	*Lachnospiraceae* spp.	TATCAGCAGGAAGATAGTGA
10	*Lachnospiraceae* spp.	AGTCAGGTACCGTCATTTTCT
11	*Lachnospiraceae* spp.	ACTGCTTTGGAAACTGCAGAT
12	*Pseudomonas* spp.	GTAGAGGGTGGTGGAATTTC
13	*Escherichia coli*	GAGCAAAGGTATTAACTTTACTC
14	Enterobacteriaceae other than *E. coli*	CGAAACTGGCAGGCTAGAGT
15	Gammaproteobacteria	CCTGGACAAAGACTGACGCT
16	*Varibaculum* spp.	TTGAGTGTAGGGGTTGATTAG
17	*Bifidobacterium longum* including subsp. *infantis*	GAGCAAGCGTGAGTAAGTTTA
18	*Bifidobacterium bifidum*	CCGAAGGCTTGCTCCCAAA
19	*Bifidobacterium breve*	CACTCAACACAAAGTGCCTTG
20	*Bifidobacterium* spp.	GCTTATTCGAAAGGTACACTC ACCCCGAAGGG
21	*Bacteroides fragilis*	GGGCGCTAGCCTAACCAG
22	*Bacteroides* spp.	ATGCATACCCGTTTGCATGTA

Targets of the probes, taken from previous paper[15].

where Y_i is the change in Z-score for the i^{th} infant ($i = 1, \ldots, n$), and X_{i,q_k} denotes the detection of the k^{th} gut microbiota spp. group ($k = 1, \ldots, 22$) at the q^{th} time point ($q = 4, 10, 30$, and 120).

We then tested β_{i_k} for significance, controlling the false discovery rate at 20% (and again at a more stringent 5%) by using a mixed directional false discovery rate method[16]. Briefly summarising the method, we defined P_{ik} as the p-value for the test:

$$H^i_{0k} : \beta_{i_k} = 0 \qquad (3.1)$$
$$H^i_{1k} : \beta_{i_k} \neq 0$$

for $i = 1, \ldots, 4$ and $k = 1, \ldots, 22$. We then treated H_{0j} as the intersection of all H^i_{0k} over i, and H_{1k} as the union of all H^i_{1k} over i. The following procedure was then undertaken:

1. The Benjamini-Hochberg method was applied at level α to test H_{0k} against H_{1k} simultaneously for $k = 1, \ldots, 22$, based on the Bonferroni pooled p-values $P_k = \min_k(P_{ik}) \times 4$

2. R denotes the total number of null hypotheses rejected.

3. All β_{i_k} were tested, and H^i_{0k} was rejected with adjusted significance level $\alpha^* = \alpha \times R/(\text{num tests}) = 0.05 \times R/(22 \times 4)$

3.3.6 Exposure patterns for an evolving gut ecosystem

It is conceivable that, in an infant, it is not the effect of the gut microbiota at a singular time point, but rather the gut ecosystem evolving over time, which influences growth. To capture this evolution, it is possible to describe an infant's exposure to gut microbiota as a pattern over time. For example, one infant's pattern could be a gut microbiota spp. that is detected at days 4, 10, and 30, then non-detected at day 120. Each combination of possible

[16]W Guo/S Sarkar/S Peddada: Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories, in: Biometrics 66 (2010), pp. 485–492.

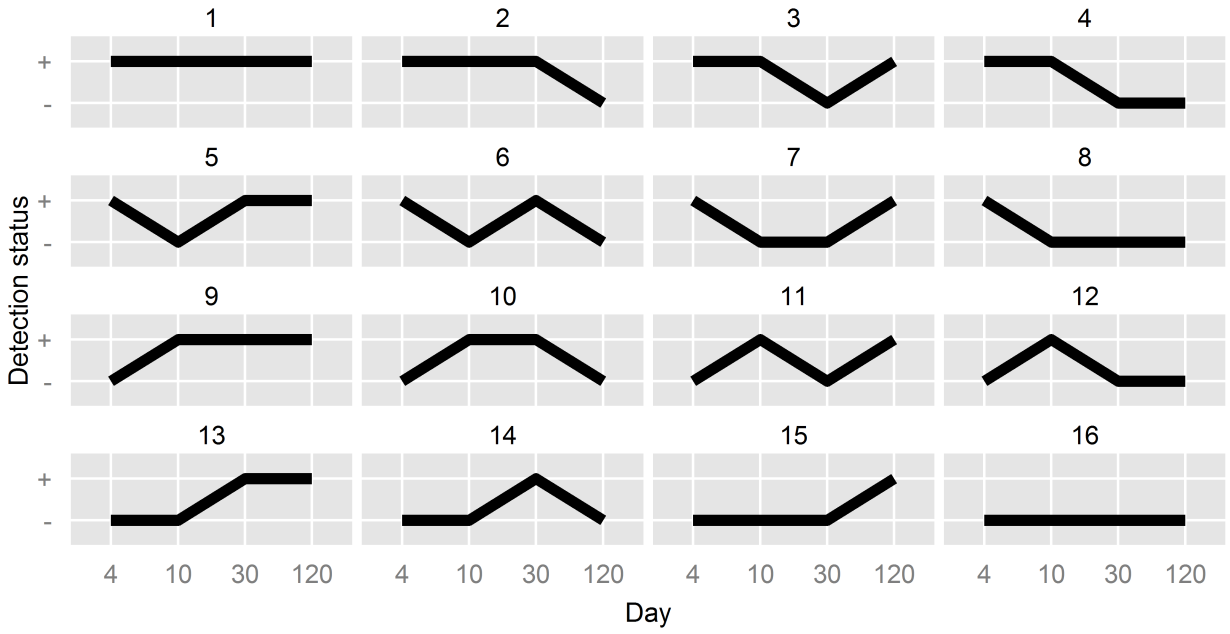

Figure 3.2. All possible exposure patterns in the data. "+" and "−" represent detection and non-detection respectively. For example, pattern 8 indicates detection at day 4, followed by non-detection at days 10, 30, and 120.

values of the gut microbiota (detected or not detected) at different time points (4, 10, 30, 120 days) was considered to be a pattern.

If a pattern was observed to occur less than 15% of the time, it was not included as a testable pattern. All 16 possible patterns are displayed in Figure 3.2. Let $\mu_{j_k,4}$ denote the population mean for the growth outcome variable (change in Z-scores, representing difference from expected growth) of infants with (four time point) pattern j, $j = 1, 2, .., 16$ for the k^{th} gut microbiota spp., $k = 1, 2, ..., 22$. Let $\hat{\mu}_{j_k,4}$ denote the estimate of $\mu_{j_k,4}$ using the sample mean and let $se(\hat{\mu}_{j_k,4})$ denote the standard error associated with the sample mean.

Using $\hat{\mu}_{j_k,4}$ and $se(\hat{\mu}_{j_k,4})$ for each pattern and gut microbiota spp. group, we applied

Tuke's method[17] to test for equivalence to zero:

$$H_0 : |\mu_{j_k,4}| \geq \epsilon,\ \epsilon > 0 \tag{3.2}$$
$$H_1 : |\mu_{j_k,4}| < \epsilon$$

where ϵ was chosen to be 0.67. This boundary was chosen because paediatricians are concerned when a child crosses two or more major centiles (2nd, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 98th) on the growth chart. For a child born on the 50th percentile (3.346 Kg boy/3.232 Kg girl) this corresponds to a change in Z-score of ± 0.67. This is a common boundary choice for studies focusing on rapid growth in infants[18]. The mean of our growth variable was chosen, instead of an analysis of contrasts using a linear regression model, because we were unable to select an appropriate reference pattern (from those displayed in Figure 3.2). In this analysis, we were concerned with identifying which gut microbiota spp. group patterns corresponded to a mean change in Z-score that was significantly close to zero (i.e. did not deviate from expected growth). This is in contrast to the previous time-specific analysis, which was focused on the relative shift in change in Z-score, when the exposure was either present of absent.

Similar to the previous analysis, we controlled the false discovery rate at 5% by using the mixed directional false discovery rate method[19]. Once we identified a significant pattern (i.e. one where $\mu_{j_k,4}$ is close to 0), we tested to see if some time points might be superfluous and not adding information; for example, it may be that only the first 30 days of exposure that affect growth, so the last time point (day 120) would not be relevant and could be removed from the pattern. From the mixed directional false discovery rate method[20], each four time

[17] J Tuke/G Glonek/P Solomon: P-values, q-values and posterior probabilities for equivalence in genomics studies, in: arVix 2012.

[18] Monteiro/Victoria: Rapid growth in infancy and childhood and obesity in later life - a systematic review (see n. 13).

[19] Guo/Sarkar/Peddada: Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories (see n. 16).

[20] Ibid.

point pattern was tested at an adjusted significance level of α^*; if the p-value of pattern $j_{k,4}$ ($P_{j_{k,4}}$) was less than $\alpha^*/2$ then (by using a Bonferroni adjustment) we had the opportunity to perform an additional test to pattern $j_{k,4}$ without risk of losing the significant result for the four time point pattern.

That is, consider the p-values of the patterns $j_{k,4}$, $j_{k,4}$ without day 120 ($j_{k,3}$), $j_{k,4}$ without days 30 and 120 ($j_{k,2}$), and $j_{k,4}$ without days 10, 30, and 120 ($j_{k,1}$), to be denoted as $P_{j_{k,4}}$, $P_{j_{k,3}}$, $P_{j_{k,2}}$, and $P_{j_{k,1}}$, respectively. The following procedures were performed after finding a four time point pattern $j_{k,4}$ whose mean is significantly close to zero:

1. If $P_{j_{k,4}} < \alpha^*/2$, then $j_{k,3}$ was tested at significance level $\alpha^*/2$
2. If $P_{j_{k,3}} < \alpha^*/3$, then $j_{k,2}$ was tested at significance level $\alpha^*/3$
3. If $P_{j_{k,2}} < \alpha^*/4$, then $j_{k,1}$ was tested at significance level $\alpha^*/4$

The process ended when a pattern's mean was either not significantly close to zero, or when $P_{j_{k,q}}$ $(q = 1, \ldots, 4)$ was not large enough to allow continued testing. This process controlled the false discovery rate, while simultaneously ensuring that no significant finding was subsequently lost by the additional testing to remove superfluous time points. A short proof, that this adaptation still retains control of the false discovery rate, is provided in the appendix. By implementing this adaptation, the resultant test was:

$$H_0 : \min\left(|\mu_{j_{k,4}}|, |\mu_{j_{k,3}}|, |\mu_{j_{k,2}}|, |\mu_{j_{k,1}}|\right) \geq 0.67$$
$$H_1 : \min\left(|\mu_{j_{k,4}}|, |\mu_{j_{k,3}}|, |\mu_{j_{k,2}}|, |\mu_{j_{k,1}}|\right) < 0.67$$

The data reduction process was only considered from the right side of the pattern to avoid confounding. By definition, a confounder must affect both the exposure and outcome, and it is not possible for an exposure at day 120 to affect the exposure between days 4 and 30. In contrast, an exposure at day 4 may influence the exposure at day 10, and is therefore a possible confounder. We stress that, by only undertaking this process on the right side of

the pattern, we do not imply that the right side of the pattern is less important. Instead, we view the process as adding information where possible (by culling superfluous points on the right side of the pattern) and leaving the pattern otherwise alone.

3.3.7 Post-hoc screening of results

If a pattern was found to have its mean significantly close to zero (i.e. the null hypothesis in (3.2) is rejected), the mean of the pattern's crude contrast (i.e. if detection at days 4 and 10 was significant, the crude contrast would be non-detection at days 4 and 10) was tested for difference to zero, using a t-test at $\alpha = 5\%$. If the crude contrast was not found to be significantly different from zero, the pattern was discarded from the significant findings. In the event of a significant crude contrast, a Welch two sample t-test was performed to test if the means of the pattern and crude contrast differed from each other. This test was performed at a significance level of $\alpha = 10\%$ due to the decrease in sample size (and hence power) when only considering the set of infants with either the pattern of interest or the crude contrast. Tests found to be significant at $\alpha = 5\%$ were noted as such.

3.4 Results and discussion

Our outcome (the difference in Z-scores of weight-for-age for 6-months versus birth) was not centred around 0 (mean/median of $-0.29/-0.38$ and $-0.13/-0.18$ for females and males respectively), which raised concerns that our sample population was inappropriate for the World Health Organization's growth curves. We investigated the larger Norwegian Human Milk Study cohort (n=3529), of which NOMIC is a subsample[21]. We found that the median weight-for-age Z-score at birth was 0.76, decreasing to 0.31 at 6 months of life, inferring a large proportion of macrosomic infants. However, if the Norwegian infants were naturally born longer, then we would expect a naturally higher birth weight; we found the median

[21] M Eggesbo et al.: Levels of hexachlorobenzene (HCB) in breastmilk in relation to birth weight in a Norwegian cohort, in: Environmental Research 109 (2009), pp. 559–66.

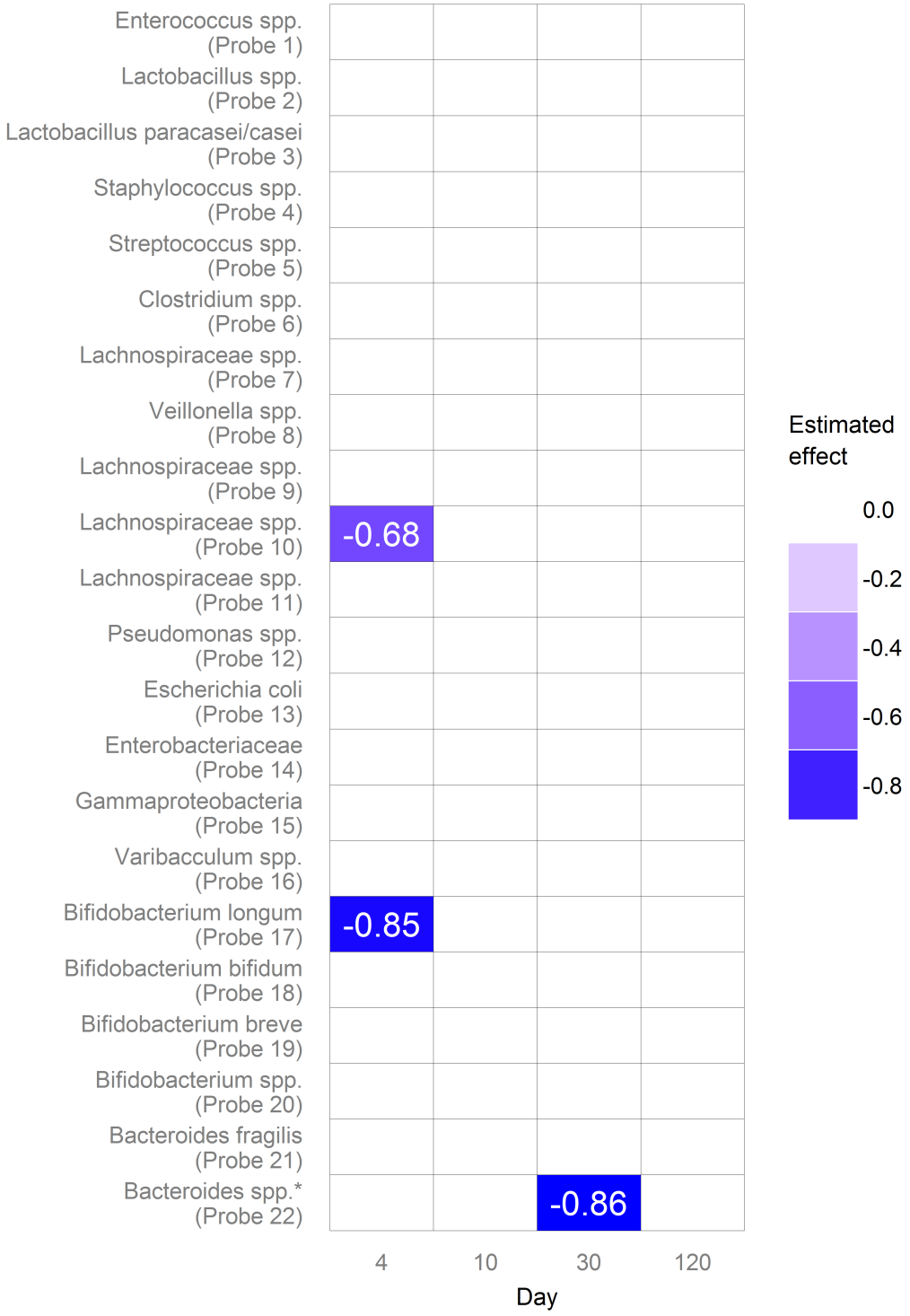

Figure 3.3. Results from the time-specific analyses for males. Coloured areas indicate significant results at 20% FDR, and are labelled with their effect estimates, while white areas indicate non-significant results. Significant results at 5% FDR are indicated by $*$. Only the results for males are displayed, as no significant results were found for females.

weight-for-length Z-score at birth was 0.63, decreasing to 0.06 at 6 months. This suggests that the Norwegian infants were born with more mass than one would expect for their appropriate length. These findings from the larger Norwegian Human Milk Study cohort were similar to what we found in NOMIC. Similar results have been shown in the Norwegian Medical Birth Registry, where it has been found that from the early 1970s to the late 1990s the birthweight of Norwegian infants has been increasing[22]. These findings strengthen the recommendations from the Norwegian Health Directorate to use the World Health Organization's growth curves[23]. It is also worth noting that because the female distribution is centred so far from zero (mean/median of $-0.29/-0.38$), we lack power when detecting gut microbiota patterns that results in a positive change in Z-score.

We applied the above methods to each gut microbiota spp. group in Table 4.2 and displayed significant time-specific results in Figure 3.3 and pattern results in Figures 3.4 and 3.5.

In the time-specific analyses, with 5% FDR, we found the detection of *Bacteroides spp.* (Probe 22) at day 30 to be significantly associated with reducing growth in males, when compared to non-detection (Figure 3.3). The current literature shows that *Bacteroides spp.* is protective against obesity[24].

In the pattern analyses, we note that the detection of *Staphylococcus* spp. (Probe 4) at day 4 was associated with expected growth in females and males (Figures 3.4 and 3.5). The literature highlights that colonisation of *Staphylococcus* spp. is a normal feature of healthy gut flora[25]. We also found that *Escherichia coli* (Probe 13) detection from day 4 through to 30 was associated with expected growth in males (Figure 3.5). The current literature indicates

[22]R Skjaerven/H Gjessing/L Bakketeig: Birthweight by gestational age in Norway, in: Acta Obstetricia et Gynecologica Scandinavica 79 (2000), pp. 440–449.

[23]Arbeidsgruppe: Nasjonale faglige retningslinjer for veiing og maaling i helsestasjons - og skolehelsetjenesten (see n. 11).

[24]G Musson/R Gambino/M Cassader: Obesity, Diabetes, and Gut Microbiota, in: Diabetes Care 33 (2010), pp. 2277–2284.

[25]P Mackowiak: The normal microbial flora, in: The New England Journal of Medicine 307 (1982), pp. 83–93.

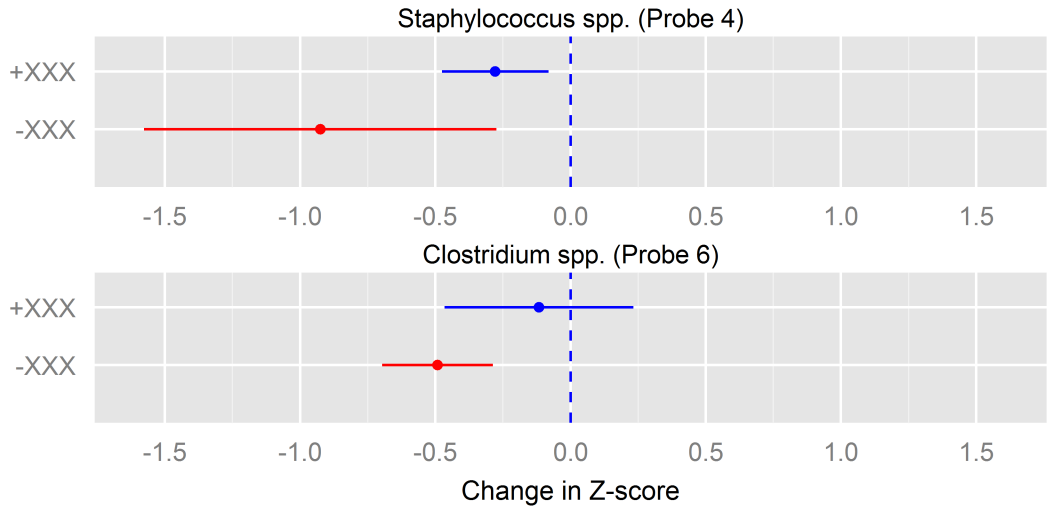

Figure 3.4. Results from the pattern analysis for females. The exposure pattern is represented by four characters, constructed from "+", "−", and "X", which represent detection, non-detection, and irrelevance, respectively, for the four time points of the analysis (days 4, 10, 30, and 120). The blue points and lines represent estimated means and 95% confidence intervals for patterns that were found to be significantly close to zero at an FDR of 5%. The crude contrasts (i.e. if "+−XX" was significant, the crude contrast would be "−+XX") that were significantly different to zero at $\alpha = 5\%$ have their estimated means and 95% confidence intervals displayed in red. For the testing of the difference of the means of the two patterns, significant results (at $\alpha = 5\%$) is indicated by $*$, otherwise significance is $\alpha = 10\%$.

that colonisation of *Escherichia coli* is a normal feature of healthy gut flora development[26].

We were concerned that our pattern analysis findings were caused by confounding that occurred before four days of life. When comparing infants with detected *Staphylococcus* spp. (Probe 4) at day 4 to those without, we found evidence that males with non-detected *Staphylococcus* spp. (Probe 4) at day 4 had lower birthweight (mean 3.19 Kg vs 3.58 Kg) and higher proportion of usage of the newborn intensive care unit, (25% vs 6%), however, these findings were inverted in the female stratum (3.68 Kg vs 3.55 Kg and 0% vs 5%), and we therefore found no conclusive evidence of confounding. We also found no noticeable

[26]HK Park et al.: Molecular analysis of colonized bacteria in a human newborn infant gut, in: The Journal of Microbiology 43 (2005), pp. 345–353.

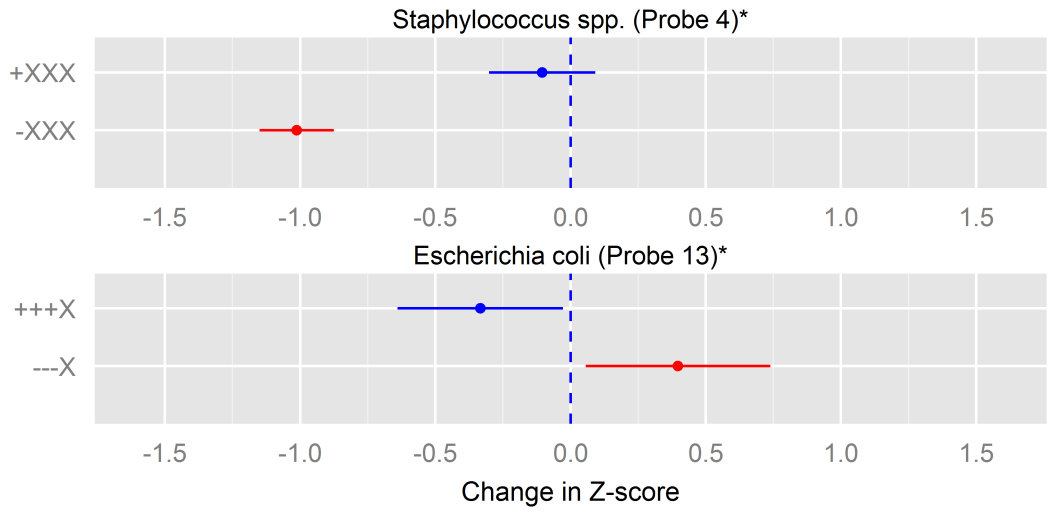

Figure 3.5. Results from the non-parametric patterns analysis for males. The exposure pattern is represented by four characters, constructed from "+", "−", and "X", which represent detection, non-detection, and irrelevance, respectively, for the four time points of the analysis (days 4, 10, 30, and 120). The blue points and lines represent estimated means and 95% confidence intervals for patterns that were found to be significantly close to zero at an FDR of 5%. The crude contrasts (i.e. if "+−XX" was significant, the crude contrast would be "−+XX") that were significantly different to zero at $\alpha = 5\%$ have their estimated means and 95% confidence intervals displayed in red. For the testing of the difference of the means of the two patterns, significant results (at $\alpha = 5\%$) is indicated by $*$, otherwise significance is $\alpha = 10\%$.

differences in the rates of preeclampsia, poor fatal growth, gestational age, or maternal BMI. No noticeable differences were found in any of the above variables when checking for confounding in *Escherichia coli* (Probe 13).

By investigating one overarching theme ("how does the gut microbiota affect infant growth?") through two different questions, we obtained two different set of results. We note that these two set of results are not mutually exclusive, nor contrasting in nature. Instead they offer different perspectives: the time-specific analysis aids in highlighting where gut microbiota has an association with the mean of the outcome, which is useful in situations where the outcome is shifted away from 0 and it is hard to find a true "healthy reference group". The pattern analysis is useful in identifying how the gut microbiota develops over time in babies with expected growth (i.e. we found that *Escherichia coli* (Probe 13) detection from day 4 through to 30 was associated with expected growth in males). This allowed us to combine a number of exposures over time, which, when viewed together, formed a cohesive message about the outcome. The message was that certain patterns corresponded to expected growth, and deviation from those patterns was associated with not achieving expected growth – instead of only identifying singular gut microbiota exposures that shifted growth.

It is important to note that as no contrasts (beyond the crude contrasts) were compared to the "expected growth" pattern, we cannot make inferences about the association between expected growth and patterns that are partially different from the "expected growth" pattern. We can only assert that the presence of particular exposure patterns are associated with expected growth, and that they significantly differed from their crude contrasts (which were also significantly different from expected growth).

When considering the application of the pattern analysis method to other analyses, it is important to note that it cannot account for confounding. We propose that in situations where confounding variables are at work, the above method be used to extract a plausible reference pattern, and then a traditional logistic regression strategy should be implemented

to address confounding. This process of reference pattern selection adds value to the current methodology literature, as it enables the transparent selection of a sensible reference pattern in scenarios (such as the one above) where it is not a simple matter to select a baseline *a priori*.

In certain situations, the outcome may be dependent on the interaction between two gut microbiota spp. groups, which would result in the above method not being appropriate without an extension. By creating patterns consisting of two – or more – gut microbiota spp. groups, and then applying the methods described here, the intra-gut microbiota spp. group dependencies can be accounted for.

As with all methods, we are limited by the granularity of our longitudinal observations and the observational nature of our data. Our method identifies time-dependent points that may contain information about potential time-dependent exposure windows that are reflected in the observed data. That is, if one assumes there is a time-dependent exposure window requiring a microbe to be detected between 100-110 days, but the microbe does not simply dissipate from the body at day 111, so a strong relationship exists between day 110 and 120, then the method will identify a time-dependent point at day 120 (reflecting the time-dependent exposure window at days 100-110). This is simply a feature of the data, and the length of time surrounding each time-dependent exposure window when it is reflected in the data (i.e. when the microbes remain similar) may vary from microbe to microbe and be dependent on the situation at hand.

The only way to prove that a time-dependent exposure window has occurred is through experiments. Using observational data, our method provides a novel way to describe potential time-dependent exposure windows that may have been reflected into the observable data. These descriptions can be further used to create time-dependent hypotheses for experiments concerned with the existence of time-dependent exposure windows. Furthermore, we highlight that our statistical methods were designed to control the false discovery rate, over a large number of tests. In doing so, it is likely that we discarded a number of clinically

significant findings that were not found to be statistically significant. We therefore make no claims about the gut microbiota spp. groups that were not found to have any significant results, as the absence of evidence is not evidence of absence.

3.5 Conclusion

Our results expand on the current literature relating gut microbiota to growth, in both methodology and biological findings. With regards to methodology, we developed a novel method to analyse longitudinal data that contains information about the development of an ecosystem over time. Crucially, this method controls the false discovery rate associated with multiple levels of multidimensional testing. We expanded the biological literature by reporting time-dependent patterns associated with expected growth, which, in some cases, confirmed the importance of gut microbiota spp. groups previously reported on.

Chapter 4

Using Bayesian distributed lag two-part models to investigate the effect of persistent organic pollutants (POPs) on gut microbiota in infants

4.1 Abstract

Gut microbiota has a critical role in human health; understanding its role in early infancy is of particular interest due to the time dependent windows that rely on microbial stimulus from the gut. That is, the development of tolerance and the optimal functioning of angiogenesis and stress responses later in life, require time dependent actions in the gut. Persistent organic pollutants (POPs) are widespread environmental contaminants that are resistant to environmental degradation through normal processes, which causes them to bioaccumulate in human and animal tissue and biomagnify in food chains. POPs are known carcinogens that disrupt natural human systems (endocrine, reproductive, and immune). Using novel statistical models, we investigated the impact of POPs (in particular, non-dioxin-like poly-

chlorinated biphenyl, IUPAC no.: 153; "PCB153") on human health through the disruption of natural gut microbiota establishment in infants. We created novel distributed lag two-part models to account for the cumulative exposure of POPs. We then identified significant associations concerning POPs affecting gut microbiota species (spp.) groups (from birth through to day 120 of life). Strong associations were found between POPs and *Bifidobacterium* spp., *Bifidobacterium bifidum*, and *Lactobacillus* spp.. Using these findings we successfully identified gut microbiota as a potential vector through which POPs may harm humans; examples were given for POPs acting as carcinogens and diarrhoeal agents.

4.2 Introduction

Persistent organic pollutants (POPs) are widespread environmental contaminants that are resistant to environmental degradation through normal processes; this causes them to bioaccumulate in human and animal tissue and biomagnify in food chains. Exposure to POPs can cause death, as well as other illnesses related to the disruption of the endocrine, reproductive, and immune systems[1]. Furthermore, POPs are known carcinogens[2]. POPs are currently regulated by the Stockholm convention (http://www.pops.int) and following the ban on them as fungicides, environmental levels have declined by more than 90% worldwide. However, POPs are still used as industrial chemicals and are created as an unintended by-product from several processes, such as production of chlorinated solvents. Therefore, population exposure to POPs is likely to continue, and as such, deserves attention.

Gut microbiota has a critical role in human health[3]; understanding its role in early infancy is of special interest because the early life period is a determinant for the subsequent adult-like microbiota. Once the first microbes arrive in the sterile gut of the newborn, a

[1]L Ritter et al.: Persistent organic pollutantsgut, in: United Nations Environment Programme 2007.

[2]B Fisher: Most unwanted, in: Environmental Health Perspectives 107 (1999), A18–A23.

[3]Backhed et al.: Host-bacterial mutualism in the human intestine (see n. 1); Mitsuoka: Intestinal flora and aging (see n. 1); Roung/Mazmanian: The gut microbiota shapes intestinal immune responses during health and disease. (See n. 1); Turnbaugh et al.: A core gut microbiome in obese and lean twins (see n. 1); Bjorksten et al.: The intestinal microflora in allergic estonian and swedish 2-year-old children (see n. 1); Mazmaniam/ Round/Kasper: A microbial symbiosis factor prevents intestinal inflammatory disease. (See n. 1).

dynamic process starts where activation of genes and expression of receptors in the host plays an important role for the building of niches and the further selection of microbes. More importantly, studies on germ free animals have revealed the presence of developmental windows that rely on microbial stimulus from the gut[4] (i.e. development of tolerance[5]), but also for optimal functioning of diverse systems, such as angiogenesis[6] and stress responses[7]. However, we still have limited knowledge of early gut microbiota[8].

We investigate the associations between POPs (in particular, non-dioxin-like polychlorinated biphenyl, IUPAC no.: 153; "PCB153") and disrupted natural gut microbiota establishment (from day 4 through to day 120), to identify a possible vector through which POPs could affect human health. Our high quality dataset includes a daily exposure profile for POPs in the infants, which allows our investigation to analyze the cumulative effect of POPs on gut microbiota. To achieve this, we developed a novel statistical method to overcome the problem of estimating the cumulative effect of POPs when large numbers of infants have outcomes of zero - non-detected gut microbiota species (spp.) groups.

[4]Thompson/Wang/Holdes: The immediate environment during postnatal development has long-term impact on gut community structure in pigs (see n. 2).

[5]Sudo et al.: The requirement of intestinal bacterial flora for the development of an ige production system fully susceptible to oral tolerance induction (see n. 3); Mazmanian et al.: An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system (see n. 3).

[6]Stappenbeck/Hooper/Gordon: Developmental regulation of intestinal angiogenesis by indigenous microbes via paneth cells (see n. 6).

[7]Sudo et al.: Postnatal microbial colonization programs the hypothalamic-pituitary-adrenal system for stress response in mice (see n. 7).

[8]M Wang et al.: T-rflp combined with principal component analysis and 16s rrna gene sequencing: an effective strategy for comparison of fecal microbiota in infants of different ages. In: Journal of Microbiological Methods 59 (2004), pp. 53–69; C Palmer et al.: Development of the human infant intestinal microbiota. In: PLoS Biology 5 (2007), pp. 1556–1573; C Favier et al.: Molecular monitoring of succession of bacterial communities in human neonates, in: Applied and Environmental Microbiology 68 (2002), pp. 219–26; Eggesbo et al.: Development of gut microbiota in infants not exposed to medical interventions (see n. 14).

4.3 Methods

4.3.1 Study population

The Norwegian Microflora Study (NoMIC) is a birth cohort established to study the establishment of gut microbiota during infancy and its consequences for child health. Participating mothers were recruited to the NoMIC study by a paediatrician at the maternity ward in a county hospital in South Norway. Recruitment was designed to purposefully oversample preterm children; whenever a preterm-birth mother was enrolled, two mothers of consecutively born term infants were recruited. The recruitment started in November 2002 and was completed in May 2005. Eligibility criteria for the study were that mothers must be fluent in Norwegian and resident in the pertinent geographic area. The study was approved by the Norwegian Data Inspectorate and the Regional Ethics Committee for Medical Research.

After the informed consent form had been signed by the mothers, containers for fecal samples and a questionnaire were provided to the participants at the maternity ward. The mothers were asked to collect and freeze one fecal sample from themselves at postpartum day 4, as well as samples from their infants when they were 4, 10, 30, and day 120 old. Study personnel retrieved the fecal samples and kept them frozen during transport to the Biobank of the Norwegian Institute of Public Health, where they were stored at -20 C upon arrival. Further questionnaires were sent to the families when their infants were aged 6, 12, 18, and 24 months.

Six hundred and one mothers agreed to participate in the NoMIC study, however, 86 (14%) of these mothers never returned any fecal samples, which left 524 infants with available fecal samples from one or more occasions. There were no checks performed on demographic data to identify if the data is not missing completely at random.

All the mothers were directed to save a 25ml milk sample from each morning for eight consecutive days, although milk sampled in a different manner was also accepted. The pooled milk samples were collected by study personnel and kept frozen during transport.

Table 4.1. Description of participants

	Subsample	Full NoMIC cohort
Sample size	73	524
Male	55.0	54.1
Maternal smokers	11.2	15.2
Siblings	0 (0, 1)	1 (0, 1)
Birth weight (Kg)	3.3 (2.7, 3.7)	3.4 (2.6, 3.8)
Maternal age	29 (27, 32)	30 (27, 33)
Maternal BMI	23.6 (21.7, 26.2)	23.9 (21.2, 27.2)

List of characteristics in the whole NoMIC study and those selected to have milk analysis. Results are presented as either % or median (interquartile range). Sample size is presented as number in selection.

The median age at start of sampling was 33 days after delivery (min 2 days, max 124; 5th percentile 16 and 95th percentile 65).

For reasons of cost, we randomly selected 73 infants from among the term children in the NoMIC cohort, and in whom milk samples had been collected, for toxicant analysis. Concentrations of non-dioxin-like polychlorinated biphenyl (ndl-PCBs; IUPAC no.: 153, "PCB153") were then measured in approximately 15 ml of breast milk at the Norwegian School of Veterinary Science in Oslo. Characteristics of the NoMIC cohort and the subsample are presented in Table 4.1.

4.3.2 Outcomes

The outcomes of interest were intensity readings for 22 probes, encoding for different gut microbiota spp. groups at 4, 10, 30, and 120 days since birth. The gut microbiota spp. groups, labelling sequence, and target bacteria (according to TNTProbeTool program) are displayed in Table 4.2.

16S rRNA gene clone libraries were constructed from DNA extracted from the fecal samples obtained on days 4, 10, 30, and 120. More detailed information about this process

can be found in a previous paper[9].

These intensity levels were analysed as a continuous variable (Y_i), which was rescaled to fall between 0 and 100, for the subset of 73 infants. Each gut microbiota spp. group was analyzed separately.

4.3.3 Exposures

The exposures of interest were the blood lipid POPs levels (PCB153) in the infants. To estimate these levels across the first day 120 of life, we used a previously developed and validated toxicokinetic model[11]. Briefly, this toxicokinetic model has two compartments representing maternal and child lipids that are connected through placental diffusion and excretion/intake of breast milk. The mother is exposed to POPs through ingestion of contaminated food whereas the child is exposed both through placental diffusion and consumption of breast-milk. Elimination of POPs from the maternal and child compartments was parameterised based on published half-life values.

We generated profiles of blood lipid POPs levels, for each child, by including information on age at delivery, pre-pregnancy weight, weight gain during pregnancy and loss thereafter, gestational age, sex, child weight at birth and onwards, duration of breast-feeding and the percentage of total intake attributable to breast-feeding when other food is incorporated into child's diet. Oral intake in mothers was optimised to match simulated breast milk levels to those measured in the study. Subsequently, every child POPs levels were estimated (for every 24 hours of life) to be used as estimates of their exposure to POPs.

To aid in interpretation of the results, the concentrations of the POPs were rescaled such that 1 unit corresponded to the difference between the 25th and 75th percentiles (of the subset of 73 infants) on day 0 of life.

[9]Eggesbo et al.: Development of gut microbiota in infants not exposed to medical interventions (see n. 14).
[11]M Verner: Submitted, in: 2012.

Table 4.2. Probes and their targets

#	Probe match	Labelling probes
1	*Enterococcus* spp.	TCATCCCTTGACGGTATCTAA
2	*Lactobacillus* spp.	GTCAAATAAAGGCCAGTTACTA
3	*Lactobacillus paracasei/casei*	CAGTTACTCTGCCGACCATT
4	*Staphylococcus* spp.	ACACATATGTTCTTCCCTAATAA
5	*Streptococcus* spp. (α-hemolytic)	AGTGTGAGAGTGGAAAGTTCA
6	*Clostridium* spp.	TCAACTTGGGTGCTGCATTC
7	*Lachnospiraceae* spp.	AGCTAGAGTGTCGGAGAGG
8	*Veillonella* spp.	GATTGGCAGTTTCCATCCCAT
9	*Lachnospiraceae* spp.	TATCAGCAGGAAGATAGTGA
10	*Lachnospiraceae* spp.	AGTCAGGTACCGTCATTTTCT
11	*Lachnospiraceae* spp.	ACTGCTTTGGAAACTGCAGAT
12	*Pseudomonas* spp.	GTAGAGGGTGGTGGAATTTC
13	*Escherichia coli*	GAGCAAAGGTATTAACTTTACTC
14	Enterobacteriaceae other than *E. coli*	CGAAACTGGCAGGCTAGAGT
15	Gammaproteobacteria	CCTGGACAAAGACTGACGCT
16	*Varibaculum* spp.	TTGAGTGTAGGGGTTGATTAG
17	*Bifidobacterium longum* including subsp *infantis*	GAGCAAGCGTGAGTAAGTTTA
18	*Bifidobacterium bifidum*	CCGAAGGCTTGCTCCCAAA
19	*Bifidobacterium breve*	CACTCAACACAAAGTGCCTTG
20	*Bifidobacterium* spp.	GCTTATTCGAAAGGTACACTCACCCCGAAGGG
21	*Bacteroides fragilis*	GGGCGCTAGCCTAACCAG
22	*Bacteroides* spp.	ATGCATACCCGTTTGCATGTA

Targets of the probes found using TNTProbeTool program, taken from previous paper[10].

4.3.4 Confounders

Information on potential confounders were obtained by questionnaires filled in by the mothers and the Medical Birth Registry of Norway.

It was decided *a priori* that potential confounders were breastfeeding, education, gender, gestational age, maternal age, parity, and smoking. All confounders were then added to the models and stepwise reduction procedures were used to identify confounders (those that affected the 95th percentile effect estimates of POPs by more than 10% in the gut microbiota spp. groups that had a possible association). These were found to be all of the variables tested, and were denoted as $\vec{Q}$.

Breastfeeding was analysed as a continuous variable at days 4, 10, and 30. The variable had four levels, corresponding to the child never having any milk substitutes, once or twice having milk substitutes, once or twice a week having milk substitutes, and only having milk substitutes. Breastfeeding was also analysed as the proportion of meals that were breast milk in the fourth month of life (corresponding to the time period close to day 120). Education was analysed as a continuous variable, with four levels corresponding to less than high school, high school, completed some of a university degree, and having completed a university degree. Gestational age was calculated based on the last menstrual period. In Norway ultrasound is routinely performed in the second trimester, and we used the ultrasound-based estimate only if the discrepancy between the two exceeded 14 days[12]. Parity was analysed as a continuous variable containing the number of siblings. Maternal smoking status at the beginning of pregnancy was dichotomised into never/previously and occasional/daily.

[12]T Henrisken et al.: Bias in studies of preterm and postterm delivery due to ultrasound assessment of gestational age. In: Epidemiology 6 (2005), pp. 533–537; B Blondel et al.: Algorithms for combining menstrual and ultrasound estimates of gestational age: consequences for rates of preterm and postterm birth. In: BJOG: An International Journal of Obstetrics and Gynaecology 109 (2002), pp. 718–720.

4.3.5 Models

Our models were designed to analyze the cumulative effect of POPs on the 22 gut microbiota spp. groups. To avoid any issues caused by infant gut microbiota's ill-defined and non-parametric relationship with time, cross sectional analyses were performed with the outcome at days 4, 10, 30, and 120. Due to the large number of non-detections ($Y_i = 0$), the data were primarily analysed using two-part models[13]. The following models were applied individually to each of the 22 gut microbiota spp. groups.

For the model with the outcome at day 120, distributed lag two-part models were implemented to estimate the cumulative effect of POPs on the intensity readings of probes corresponding to gut microbiota spp. groups[14]:

$$E[Y_{(120)i}|\vec{Q}_{(t120i}, \vec{X}_{(120)i}] = P(Y_{(120)i} \geq 0|\vec{Q}_{(120)i}, \vec{X}_{(120)i}) \times E[Y_{(120)i}|Y_{(120)i} \geq 0, \vec{Q}_{(120)i}, \vec{X}_{(120)i}] \tag{4.1}$$

where infants $i = 1, \ldots, 73$, $\vec{Q}_{(120)i}$ were the confounders, $\vec{X}_{(120)i}$ was a vector containing blood concentrations of POPs at days $0, 10, 20, \ldots,$ and 120.

$P(Y_{(120)i} \geq 0|\vec{Q}_{(120)i}, \vec{X}_{(120)i})$ was estimated using a distributed lag probit regression:

$$P(Y_{(120)i} \geq 0|\vec{Q}_{(120)i}, \vec{X}_{(120)i}) = \Phi\left(\beta_0 + \vec{\beta_1}'\vec{Q}_{(120)i} + \beta_2 X_{(0)i} + \beta_3 X_{(10)i} + \beta_4 X_{(20)i} + \cdots + \beta_{14} X_{(120)i}\right)$$

and $E[Y_{(120)i}|Y_{(120)i} \geq 0, \vec{Q}_{(120)i}, \vec{X}_{(120)i}]$ was estimated using a distributed lag linear regression restricted to the detectable (non-zero) data:

$$E[Y_{(120)i}|Y_{(120)i} \geq 0, \vec{Q}_{(120)i}, \vec{X}_{(120)i}] = \beta_0 + \vec{\beta_1}'\vec{Q}_{(120)i} + \beta_2 X_{(0)i} + \beta_3 X_{(10)i} + \beta_4 X_{(20)i} + \cdots + \beta_{14} X_{(120)i}$$

Further detailed information about the construction of two-part models and distributed lag

[13] N Duan et al.: A Comparison of Alternative Models for the Demand for Medical Care. Santa Monica 1982.

[14] L Welty et al.: Bayesian distributed lag models: Estimating effects of particulate matter air pollution on daily mortality. In: Biometrics 65 (2009), pp. 282–291.

models can be found in the appendix.

For the models with outcomes at days 4, 10, and 30, distributed lag models were not needed as there were negligible differences in blood concentration from day 0 through to 30. Hence only the concentrations of POPs at the day of the outcome were used as the measure of exposure (i.e cumulative exposure contained the same information as acute exposure):

$$E[Y_{(t)i}|\vec{Q}_{(t)i}, X_{(t)i}] = P(Y_{(t)i} \geq 0|\vec{Q}_{(t)i}, X_{(t)i}) \times E[Y_{(t)i}|Y_{(t)i} \geq 0, \vec{Q}_{(t)i}, X_{(t)i}]$$

where infants $i = 1, \ldots, 73$, $\vec{Q}_{(t)i}$ were the confounders, $X_{(t)i}$ were the blood concentrations of POPs at time (t), and (t) identifies which of the three cross sectional models is being referred to $(t = 4, 10, 30)$. $P(Y_{(t)i} \geq 0|\vec{Q}_{(t)i}, X_{(t)i})$ was calculated using a probit regression, while $E[Y_{(t)i}|Y_{(t)i} \geq 0, \vec{Q}_{(t)i}, \vec{X}_{(t)i}]$ was calculated using a linear regression restricted to the detectable (non-zero) data.

When there were less than five non-detections, the data were analysed using linear regression models. Further detailed information regarding the construction of these models is available in the appendix.

4.4 Results and discussion

Our time-specific findings for spp. groups, which have at least one significant result, are displayed in Table 4.3. From Table 4.3 we can see that *Bifidobacterium* spp. (Probe 20), *Bifidobacterium bifidum* (Probe 18), and *Lactobacillus* spp. (Probe 2) have strong, clearly defined associations with POPs.

Bifidobacterium spp. (Probe 18) had a significant negative association with POPs at all four time points, which was increasing in strength over time (Table 4.3). Some *Bifidobacterium* strains have been shown to produce folate in the colon[15] in addition to transport

[15] A Pompei/et al: Folate production by bifidobacteria as a potential probiotic property. In: Applied and Environmental Microbiology 73 (2007), pp. 179–185.

Table 4.3. Significant associations

#	Probe match	Day 4	Day 10	Day 30	Day 120
1	*Enterococcus* spp.	-7.6 $(-13.4, 4.4)$	-11.4^* $(-16.4, -3.1)$	-2.3 $(-6.9, 5.7)$	-14.6^* $(-29.3, -2.8)$
2	*Lactobacillus* spp.	-5.8^* $(-12.1, -0.3)$	-4.7 $(-12.4, 3.8)$	1.4 $(-5.0, 13.1)$	1.8 $(-4.1, 11.9)$
18	*Bifidobacterium bifidum*	-9.5^* $(-16.4, -2.9)$	-9.9^* $(-17.1, -2.5)$	-9.0^* $(-13.8, -2.9)$	-13.7^* $(-28.4, -2.3)$
20	*Bifidobacterium* spp.	-11.5^* $(-17.7, -3.5)$	-6.2^* $(-11.8, -0.3)$	-6.2 $(-11.0, 0.0)$	-3.6 $(-10.3, 4.0)$
3	*Lactobacillus paracasei/casei*	-3.1 $(-15.3, 3.5)$	-6.6^* $(-14.1, -1.5)$	-1.7 $(-6.6, 6.7)$	-5.6 $(-17.7, 4.8)$
4	*Staphylococcus* spp.	-3.1 $(-11.6, 5.6)$	1.9 $(-5.1, 9.0)$	-2.2 $((-12.6, 4.4))$	-5.8^* $(-10.2, -1.3)$

Probes with at least one significant association found. Significance at $\alpha = 5\%$ is denoted by *

anticancer genes into tumours[16]. We note that POPs have been shown to play a role in cancer causation[17].

In contrast to the previous sustained association, *Bifidobacterium bifidum* (Probe 18) only had a significant negative association with POPs at days 4 and 10 (borderline significant at day 30), with the association decreasing in magnitude over time (Table 4.3). *Bifidobacterium bifidum* has been shown to be protective against diarrhoeal diseases[18] and allergies[19], while POPs are known to negatively impact immune systems[20].

Finally, we found that *Lactobacillus* spp. (Probe 2) had a similar significant negative association with POPs at day 4, decreasing to a slightly smaller in magnitude (and non-significant) association at day 10 (Table 4.3). As with *Bifidobacterium*, some *Lactobacillus*

[16]X Li/et al: Bifidobacterium adolescentis as a delivery system of endostatin for cancer gene therapy: Selective inhibitor of angiogenesis and hypoxic tumor growth. In: Cancer Gene Therapy 10 (2003), pp. 105–111.

[17]Fisher: Most unwanted (see n. 2).

[18]J Saavedra et al.: Feeding of bifidobacterium bifidum and streptococcus thermophilus of inants in hospital for prevention of diarrhoea and shedding of rotavirus. In: The Lancet 344 (1994), pp. 1046–1049.

[19]J Kim et al.: Effect of probiotic mix (bifidobacterium bifidum, bifidobacterium lactis, lactobacillus acidophilus) in the primary prevention of eczema: a double-blind, randomized, placebo-controlled trial. In: 21 (2010), pp. 386–393.

[20]Ritter et al.: Persistent organic pollutantsgut (see n. 1).

strains have been shown to act as anticancer agents[21], and again noting that POPs have been shown to play a role in cancer causation[22].

We developed and implemented a distributed lag probit regression model. This model was then used to create a distributed lag two-part model, which was ultimately used to account for the large number of non-detected gut microbiota intensity readings. This novel distributed lag two-part model was applied to gut microbe data; while distributed lag models are common in environmental epidemiology, as far as we know, this is the first time such a method has been used in microbe data.

To the best of our knowledge, this is the first study to examine POPs affecting gut microbiota. Through the identification of such associations, we can better understand the kinetics of disease associated with POPs. Such pathways have been tentatively identified for cancer and diarrhoeal diseases caused by POPs.

[21]S Choi et al.: Effects of lactobacillus strains on cancer cell proliferation and oxidative stress in vitro, in: 42 (2006), pp. 452–458.

[22]Fisher: Most unwanted (see n. 2).

Appendix A

Novel developmental analyses identify longitudinal patterns of early gut microbiota that affect infant growth

For $k = 1, \ldots, 22$ gut microbiota spp. groups with $j = 1, \ldots, 16$ patterns, let the p-values of the patterns $j_{k,4}$, $j_{k,4}$ without day 120 ($j_{k,3}$), $j_{k,4}$ without days 30 and 120 ($j_{k,2}$), and $j_{k,4}$ without days 10, 30, and 120 ($j_{k,1}$), be denoted as $P_{j_{k,4}}$, $P_{j_{k,3}}$, $P_{j_{k,2}}$, and $P_{j_{k,1}}$, respectively. These patterns are considered to be part of the pattern family j_k. We aim to test the following:

$$H_0 : \min\left(|\mu_{j_{k,4}}|, |\mu_{j_{k,3}}|, |\mu_{j_{k,2}}|, |\mu_{j_{k,1}}|\right) \geq 0.67$$
$$H_1 : \min\left(|\mu_{j_{k,4}}|, |\mu_{j_{k,3}}|, |\mu_{j_{k,2}}|, |\mu_{j_{k,1}}|\right) < 0.67$$

Given an arbitrary significance level α^*, the following procedures are performed after finding a four time point pattern $j_{k,4}$ whose mean is significantly close to zero:

1. If $P_{j_{k,4}} < \alpha^*/2$, then $j_{k,3}$ is tested at significance level $\alpha^*/2$
2. If $P_{j_{k,3}} < \alpha^*/3$, then $j_{k,2}$ is tested at significance level $\alpha^*/3$

3. If $P_{j_{k,2}} < \alpha^*/4$, then $j_{k,1}$ is tested at significance level $\alpha^*/4$

The Bonferroni adjusted p-value for gut microbiota spp. group k is denoted as

$$P_k = (\text{num patterns tested in gut microbiota spp. group } k) \times min_{j \in k}(P_{j_{k,4}})$$

Without loss of generality, we will prove that the FDR is controlled when step 1 is implemented - proofs for the other steps are similarly formed. This proof relies on the proof provided in Guo et al. (2010). We let V be the number of false positives, $I_1 = \{1 \leq k \leq 22 | H_1\}$ be the set of indices of false null hypotheses, $\alpha^* = \alpha \times R/(22 \times 16)$ (obtained from Guo et al., 2010), and $I(\cdot)$ be an indicator function. We then express V as:

$$\begin{aligned} V &= \sum_{k \in I_1} I\left(\cup_{j=1}^{16} (\text{Reject } H_0 \text{ for } j_k | H_0)\right) \\ &= \sum_{k \in I_1} I\left(\cup_{j=1}^{16} \left(\left(\alpha^*/2 \leq P_{j_{k,4}} \leq \alpha^* | H_0\right) + \left(P_{j_{k,4}} \leq \alpha^*/2, P_{j_{k,3}} \leq \alpha^*/2 | H_0\right)\right)\right) \end{aligned}$$

Then

$$\begin{aligned} FDR =& E\left\{\frac{V}{R \vee 1}\right\} \\ =& E\left\{\frac{E(V|R=r)}{R \vee 1}\right\} \\ =& E\left[\frac{\sum_{j \in I_1} \Pr\left\{\cup_{j=1}^{16} \left(\left(\alpha^*/2 \leq P_{j_{k,4}} \leq \alpha^* | H_0\right) + \left(P_{j_{k,4}} \leq \alpha^*/2, P_{j_{k,3}} \leq \alpha^*/2 | H_0\right)\right) | R=r\right\}}{R \vee 1}\right] \\ =& \sum_{r=1}^{22} \sum_{k \in I_1} \frac{1}{r} \Pr\left\{\cup_{j=1}^{16} \left(\left(\alpha^*/2 \leq P_{j_{k,4}} \leq \alpha^*, R=r | H_0\right) + \left(P_{j_{k,4}} \leq \alpha^*/2, P_{j_{k,3}} \leq \alpha^*/2, R=r | H_0\right)\right)\right\} \\ \leq& \sum_{r=1}^{22} \sum_{k \in I_1} \sum_{j=1}^{16} \frac{1}{r} \left\{\Pr\left(\alpha^*/2 \leq P_{j_{k,4}} \leq \alpha^*, R=r | H_0\right) + \Pr\left(P_{j_{k,4}} \leq \alpha^*/2, P_{j_{k,3}} \leq \alpha^*/2, R=r | H_0\right)\right\} \end{aligned}$$

The above inequality follows from the Bonferroni inequality.

This proof continues on the next page.

$$\leq \sum_{r=1}^{22} \sum_{k \in I_1} \sum_{j=1}^{16} \frac{1}{r} \left\{ \Pr\left(\alpha^*/2 \leq P_{j_{k,4}} \leq \alpha^*, R = r | H_0\right) + \right.$$

$$\left. max\left(\Pr\left(P_{j_{k,4}} \leq \alpha^*/2, R = r | H_0\right), \Pr\left(P_{j_{k,3}} \leq \alpha^*/2, R = r | H_0\right)\right) \right\}$$

$$= \sum_{r=1}^{22} \sum_{k \in I_1} \sum_{j=1}^{16} \frac{1}{r} \left\{ \Pr\left(\alpha^*/2 \leq P_{j_{k,4}} \leq \alpha^* | H_0\right) \times \Pr\left(R^{(-k)} = r-1\right) + \right.$$

$$\left. max\left(\Pr\left(P_{j_{k,4}} \leq \alpha^*/2 | H_0\right), \Pr\left(P_{j_{k,3}} \leq \alpha^*/2 | H_0\right)\right) \times \Pr\left(R^{(-k)} = r-1\right) \right\}$$

The above simplification results from the assumption that each gut microbiota spp. group is independent of each other, where $R^{(-k)}$ denotes the number of rejections in the step up procedure with critical constants $a_l = \frac{l+1}{22}, l = 1, \ldots, 22-1$ based on $\{P_1, \ldots, P_{22}\} \backslash \{P_k\}$

$$\leq \sum_{r=1}^{22} \sum_{k \in I_1} \sum_{j=1}^{16} \frac{1}{r} \left\{ \alpha^*/2 \times \Pr\left(R^{(-k)} = r-1\right) + max\left(\alpha^*/2, \alpha^*/2\right) \times \Pr\left(R^{(-k)} = r-1\right) \right\}$$

The above inequality follows from $\Pr(pvalue \leq p) \leq p,$ for any $p \in (0,1)$ under H_0

$$= \sum_{r=1}^{22} \sum_{k \in I_1} \sum_{j=1}^{16} \frac{1}{r} \left\{ \alpha^* \times \Pr\left(R^{(-k)} = r-1\right) \right\}$$

$$= \sum_{r=1}^{22} \sum_{k \in I_1} \sum_{j=1}^{16} \frac{1}{r} \left\{ \alpha \times R/(22 \times 16) \times \Pr\left(R^{(-k)} = r-1\right) \right\}$$

$$= \frac{m_1}{22} \alpha$$

$$= \frac{m_1}{m} \alpha$$

$$\leq \alpha$$

Table A.1. Probes and the frequency of their detection

#	Female Day 4	Female Day 10	Female Day 30	Female Day 120	Male Day 4	Male Day 10	Male Day 30	Male Day 120
1	56	51	54	66	64	51	54	62
2	40	47	50	36	25	37	44	34
3	49	59	60	61	62	55	61	72
4	91	98	84	60	96	94	83	58
5	46	54	48	25	49	45	54	36
6	42	42	43	55	40	40	48	69
7	27	26	34	51	29	28	26	54
8	30	52	68	67	28	48	71	66
9	48	49	53	67	44	48	56	70
10	48	49	59	75	53	60	62	72
11	63	57	65	61	60	64	59	68
12	68	71	70	57	71	72	66	64
13	62	63	60	80	63	67	66	72
14	75	81	81	88	86	89	82	89
15	86	91	94	98	92	95	91	94
16	37	29	36	39	31	31	27	51
17	84	81	80	84	84	84	83	87
18	82	79	83	88	73	81	83	86
19	50	55	59	63	39	48	54	66
20	79	80	84	83	74	76	77	79
21	35	30	35	31	29	27	30	19
22	68	47	63	61	52	47	48	45

The frequency of the detection of each probe, stratified by sex and day. Information is presented as percent detected.

Appendix B

Using Bayesian distributed lag two-part models to investigate the effect of persistent organic pollutants (POPs) on gut microbiota in infants

B.1 Two-part models

The two-part model expressed in Equation 4.1 is comprised of two independent parts: $P(Y_{(t)i} \geq 0)$ and $E[Y_{(t)i}|Y_{(t)i} \geq 0, \vec{Q}_{(t)i}, \vec{X}_{(t)i}]$[1].

To estimate $P(Y_{(t)i} \geq 0)$, the intensity levels were dichotomised into detected ($Z_{(t)i} = 1$) and non-detected ($Z_{(t)i} = 0$). We model $Z_{(t)i}$ using probit regressions by introducing a latent unobserved variable ${Z^*_{(t)i}}$[2]:

$$Z_{(t)i} = \begin{cases} 1 & \text{if } Z^*_{(t)i} \geq 0 \\ 0 & \text{if } Z^*_{(t)i} < 0 \end{cases}$$

[1] Duan et al.: A Comparison of Alternative Models for the Demand for Medical Care. (See n. 13).

[2] J Albert/S Chibl: Bayesian analysis of binary and polychromous response data. In: 88 (1993), pp. 669–679.

Where $Z^*_{(t)i} \sim N\left(E[Z^*_{(t)i}], 1\right)$. This can then be interpreted as a normal regression problem where the response is in the form of grouped data[3]. We define the conditional expectation of $Z^*_{(t)i}$ as:

$$E\left(Z^*_{(t)i}|\vec{Q}_{(t)i}, \vec{X}_{(t)i}\right) = \vec{\beta}^*_{(t)}\vec{Q}_{(t)i} + \vec{\theta}^*_{(t)}\vec{X}_{(t)i}$$

where people $i = 1, \ldots, n$, $\vec{Q}_{(t)i}$ were the confounders, $\vec{X}_{(t)i}$ were the blood concentrations of POPs, and separate analyses were performed for days $t = 4, 10, 30,$ and 120.

The fully conditional posterior distributions of $Z^*_{(t)1}, \ldots, Z^*_{(t)N}$ are independent with[4]:

$$Z^*_{(t)i}|\vec{Q}_i, \vec{X}_i, Z_{(t)i} \sim \begin{cases} N(E[Z^*_{(t)i}|\vec{Q}_i, \vec{X}_i], 1) & \text{truncated at the left by 0 if } Z_{(t)i} = 1 \\ N(E[Z^*_{(t)i}|\vec{Q}_i, \vec{X}_i], 1) & \text{truncated at the right by 0 if } Z_{(t)i} = 0 \end{cases} \tag{B.1}$$

To estimate $E[Y_{(t)i}|Y_{(t)i} \geq 0, \vec{Q}_{(t)i}, \vec{X}_{(t)i}]$, we model $Y_{(t)i}$ using linear regressions on the non-zero subset of the data:

$$Y_{(t)i}|Y_{(t)i} \geq 0, Q_{(t)i}, X_{P(t)i} = \vec{\beta}_{(t)}\vec{Q}_{(t)i} + \vec{\theta}_{(t)}\vec{X}_{(t)i} + \epsilon_{(t)i}$$

where people $i = 1, \ldots, n$, $\vec{Q}_{(t)i}$ were the confounders, $\vec{X}_{(t)i}$ were the blood concentrations of POPs, and separate analyses were performed for days $t = 4, 10, 30,$ and 120.

As there were negligible differences in blood concentration from day 0 to 30, concentrations at the day of response were used as the measure of exposure to chemicals. This was not appropriate when considering an outcome at day 120, as the blood concentrations varied a non-nominal amount between day 0 and 120. To capture and account for these differences over time, a vector containing blood concentrations at days $0, 10, 20, \ldots, 110$, and 120 was used as the exposure. The models with an outcome at day 120 used distributed lag models to account for the inherent problems with fitting models containing highly correlated covariates.

[3]Albert/Chibl: Bayesian analysis of binary and polychromous response data. (See n. 2).
[4]Ibid.

Once estimated distributions were obtained for $\vec{\beta}^*_{(t)}$, $\vec{\theta}^*_{(t)}$, $\vec{\beta}_{(t)}$ and $\vec{\theta}_{(t)}$, we simulated 10,000 possible realisations of the coefficient vectors. These vectors were multiplied against $\vec{Q}_{(t)i}$, $\vec{X}_{(t)i}$, and $\vec{X}^*_{(t)i} + \vec{1}$ to estimate $E[Y_{(t)i}|\vec{Q}_{(t)i}, \vec{X}_{(t)i}]$ and $E[Y_{(t)i}|\vec{Q}_{(t)i}, \vec{X}_{(t)i} + \vec{1}]$.

The overall estimate of the effect from a 1 unit change of POPs was calculated by:

$$\Delta_{(t)i} = E[Y_{(t)i}|\vec{Q}_{(t)i}, \vec{X}_{(t)i} + \vec{1}] - E[Y_{(t)i}|\vec{Q}_{(t)i}, \vec{X}_{(t)i}]$$

$\bar{\Delta}_{(t)} = \sum_i \Delta_{(t)i}/n$ was calculated for each of the $10,000$ possible realisations of the coefficient vectors. The resultant empirical distributions of $\bar{\Delta}_{(t)}$ were then used to calculate median estimates and confidence intervals for each of the fitted models.

B.2 Distributed lag models from Welty et al (2009)

Distributed lag models are regression models that include lagged exposure variables (or distributed lags) as covariates[5]. We implemented them to estimate the short-term cumulative effect of chemical toxicants in infants on gut microbiota spp. groups.

We have outcome variables $Y_{(t)1}, \ldots, Y_{(t)N}$ (N people at time t). We aim to model these outcomes using exposure time series $X_{(t)i}$. We consider the general normal linear model:

$$E[Y_{(t)i}|X_{(1)i}, \ldots, X_{(t)i}] = \sum_{l=0}^{L} \theta_l X_{(t-l)i} \tag{B.2}$$

with $Y_{(t)i} \sim N(E[Y_{(t)i}], \sigma^2)$.

The main difficulty in distributed lag models is specifying a prior on $\vec{\theta} = (\theta_0, \theta_1, \ldots \theta_L)^T$ that is uninformative on the distributed lag coefficients for small l (i.e. close to the time of the outcome) but that constrain the coefficients with larger l (further in the past) to be smoother and approach zero.

It is assumed that $\vec{\theta} \sim N(0, \mathbf{\Omega})$, where $\mathbf{\Omega}$ is constructed so that for increasing lag the

[5]Welty et al.: Bayesian distributed lag models: Estimating effects of particulate matter air pollution on daily mortality. (See n. 14).

diagonal elements decrease to zero ($\text{Var}(\theta_l) \to 0$) and the off-diagonal elements in its correlation matrix increase to one ($\text{Cor}(\theta_{l-1}, \theta_l) \to 1$). One approach is to define $\boldsymbol{\Omega} = \boldsymbol{ABA}$, where $\boldsymbol{AA}^T$ is the diagonal matrix of the individual variances of the θ_ls and $\boldsymbol{B}$ is the correlation matrix for $\vec{\theta}$. It is possible to achieve an appropriate $\boldsymbol{\Omega}$ by setting $\boldsymbol{A}$ equal to the Cholesky decomposition of a diagonal matrix with the desired prior variances and setting $\boldsymbol{B}$ equal to the correlation matrix for increasingly correlated normal random variables.

We define:

$$\boldsymbol{V}(\eta_j) = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & \exp(\eta_j \times 1)^{1/2} & 0 & \dots & 0 \\ 0 & 0 & \exp(\eta_j \times 2)^{1/2} & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \exp(\eta_j \times L)^{1/2} \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{V}(\eta_2)\vec{F}_1 + \{\boldsymbol{I}_{L+1} - \boldsymbol{V}(\eta_2)\}\vec{1}_{L+1}F_2, \text{ where} \\ \vec{F}_1 &\sim N(0, \boldsymbol{I}_{L+1}) \\ F_2 &\sim N(0,1), \text{ which gives} \\ \boldsymbol{M}(\eta_2) &= \text{cov}(\boldsymbol{Q}) \\ &= \boldsymbol{V}(\eta_2)\boldsymbol{V}(\eta_2)^T + \{\boldsymbol{I}_{L+1} - \boldsymbol{V}(\eta_2)\}\vec{1}_{L+1} \times \vec{1}_{L+1}^T\{\boldsymbol{I}_{L+1} - \boldsymbol{V}(\eta_2)\}^T \end{aligned}$$

We then additionally define σ^2 as the prior variance of θ_0 and the hyper parameter η_1 ($\eta_1 \leq 0$) determines how quickly the prior variances of the θ_ls tend towards zero. $\vec{1}_{L+1}$ is a $(L+1) \times 1$ vector of ones and $\boldsymbol{I}_{L+1}$ is the $(L+1) \times (L+1)$ identity matrix. We then define $\boldsymbol{A} = \sigma\boldsymbol{V}(\eta_1)$ and $\boldsymbol{B} = \boldsymbol{W}(\eta_2)$; the latter of which is equal to the correlation matrix derived from the covariance matrix $\boldsymbol{M}(\eta_2)$.

$\boldsymbol{B} = \boldsymbol{W}(\eta_2)$ is the correlation matrix for the mixture of normal random variables $\boldsymbol{Q}$. The first few elements of the independent $\vec{F}_1$ are weighted more heavily than the corresponding

first few elements of the dependent $\vec{1}_{L+1}F_2$, and the latter elements of the dependent $\vec{1}_{L+1}F_2$ are weighted more heavily than the latter elements of the independent $\vec{F_1}$. The parameter η_2 controls how quickly the mixture moves from independent to dependent. Hence the prior prior on $\vec{\beta}$ is $N(0, \sigma^2\mathbf{\Omega}(\vec{\eta}))$, where $\mathbf{\Omega}(\vec{\eta}) = \boldsymbol{V}(\eta_1)\boldsymbol{W}(\eta_2)\boldsymbol{V}(\eta_1)$ and $\vec{\eta} = (\eta_1, \eta_2)^T$.

We define $\hat{\vec{\beta}}$ and $\hat{\mathbf{\Sigma}}$ as the maximum likelihood estimates of the unconstrained distributed lag coefficients and sample covariance matrix (from Equation B.2), resulting in the posterior for $\vec{\beta}$ conditional on $\vec{\eta}$ and σ:

$$\vec{\beta}|\hat{\vec{\beta}}, \vec{\eta}, \sigma^2 \sim N\left(\{1/\sigma^2\mathbf{\Omega}(\boldsymbol{\eta})^{-1} + \hat{\mathbf{\Sigma}}^{-1}\}^{-1}\hat{\mathbf{\Sigma}}^{-1}\hat{\vec{\beta}}, \{1/\sigma^2\mathbf{\Omega}(\vec{\eta})^{-1} + \hat{\mathbf{\Sigma}}^{-1}\}^{-1}\right) \tag{B.3}$$

The hyperparameter σ^2 is assumed to be ten times the estimated statistical variance of θ_0; diffuse enough to have little influence on the first few θ coefficients. After setting a discrete uniform prior of $\vec{N_1} \times \vec{N_2}$ on $\vec{\eta}$ it is possible to obtain the posterior in a closed form for a given $\vec{\eta}^*$:

$$p(\vec{\eta}^*|\hat{\vec{\theta}}) = \frac{|\sigma^2\mathbf{\Omega}(\vec{\eta}^*)\hat{\mathbf{\Sigma}}^{-1} + \boldsymbol{I}|^{-1/2}\exp\left[-\frac{1}{2}\left\{\hat{\mathbf{\Sigma}}^{-1} - \hat{\mathbf{\Sigma}}^{-1}\left(\hat{\mathbf{\Sigma}}^{-1} + \frac{1}{\sigma^2}\mathbf{\Omega}(\vec{\eta}^*)^{-1}\right)^{-1}\hat{\mathbf{\Sigma}}^{-1}\right\}\hat{\vec{\theta}}\right]}{\sum_{\eta\in\vec{N_1}\times\vec{N_2}}|\sigma^2\mathbf{\Omega}(\vec{\eta})\hat{\mathbf{\Sigma}}^{-1} + \boldsymbol{I}|^{-1/2}\exp\left[-\frac{1}{2}\left\{\hat{\mathbf{\Sigma}}^{-1} - \hat{\mathbf{\Sigma}}^{-1}\left(\hat{\mathbf{\Sigma}}^{-1} + \frac{1}{\sigma^2}\mathbf{\Omega}(\vec{\eta})^{-1}\right)^{-1}\hat{\mathbf{\Sigma}}^{-1}\right\}\hat{\vec{\theta}}\right]} \tag{B.4}$$

With sufficiently large ranges for $\vec{N_1}$ and $\vec{N_2}$, the data drives the strength (or weakness) of the prior distribution, which thus controls the eventual smoothness of the estimated function - estimated from the data.

B.3 Final Gibbs sampler implementation

We can use the equations[6] from Welty et al (2009), which are listed in the previous section, to constrain our estimates and hence adjust for the high levels of collinearity. For linear

[6]Welty et al.: Bayesian distributed lag models: Estimating effects of particulate matter air pollution on daily mortality. (See n. 14).

regressions, a standard regression is fit to obtain $\vec{\theta}_P$, $\vec{\theta}_H$, $\vec{\theta}_D$ and their respective covariance matrices. For each of the $\vec{\theta}$ and Σ, we apply Equations B.3 and B.4 in a Gibbs sampler, which constrains the parameters in the aforementioned manner. For probit models, a Gibbs sampler must be used to alternate between Equations B.1, B.3, and B.4.

To include multiple lagged variables or other variables (confounders), we make a simple alteration to the prior covariance matrix. We being by defining $\mathbf{\Sigma}_{0j} = \mathbf{\Omega}(\vec{\eta}^*)$ where $j = 1, \ldots, J$ is the index of lagged variables. We then define the prior covariance matrix as:

$$\mathbf{\Sigma}_0 = \begin{bmatrix} \mathbf{\Sigma}_{01} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{02} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & \ldots & \ldots & \mathbf{\Sigma}_{0J} & \mathbf{0} \\ 0 & \ldots & \ldots & \mathbf{0} & \psi \times \boldsymbol{I} \end{bmatrix}$$

where $\boldsymbol{I}$ is an identity matrix and ψ is a sufficiently large number to produce a non-informative variance prior on the confounders in the model, which are positioned in the bottom right quadrant of the design matrix. $\mathbf{\Sigma}_0$ can then be used to replace $\mathbf{\Omega}(\vec{\eta}^*)$ in Equation B.3. In doing so, we have created distributed lag models that allow multiple lagged variables, as well as other coefficients, which are implemented in both linear and probit regressions.